IN UTERO EXPOSURE TO MATERNAL IMMUNE ACTIVATION AND AUTISM SPECTRUM DISORDER

by
Martha Brucato

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

April, 2017

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social interaction and communication, and repetitive behavior and stereotypical interests. First we describe what is known about ASD risk factors, including genetic variants and environmental exposures, particularly during gestation (**Chapter 1**). Then we tested the hypothesis that prenatal exposure to maternal immune activation (MIA) increases the risk of ASD. In a prospective birth cohort (Boston Birth Cohort, BBC), we found that prenatal exposure to maternal fever, and not maternal genitourinary infections or influenza, is associated with an increased risk of ASD (**Chapter 2**). Electronic medical records (EMR) were used to identify children in the BBC with ASD or typical development. While reliance on EMR enables us to increase sample size compared to a traditional study that requires extensive research contact, it could lead to outcome misclassification. Here, we explored using Random Forests, a data mining and machine learning technique, and Latent Class Analysis, a probabilistic clustering method, to identify other EMR diagnosis codes that help predict a child's ASD status (**Chapter 3**). These techniques were able to identify children with typical and atypical development in the BBC.

Finally, to further explore the potential biological consequences of prenatal exposure to MIA, we analyzed DNA methylation in the whole blood of 2-5 year old children in the Study to Explore Early Development (**Chapter 4**). We found one site in an intergenic region that was differentially methylated in children whose mothers contracted

an infection shortly before they were conceived, and two sites in the genome (*IQSEC1*, *EPS8L3*) that were differentially methylated in children whose mother had an infection during her third trimester. While the differences in percent methylation were small in magnitude (<1% mean or median absolute difference), they were statistically significant after accounting for technical and biological sources of variation, including ancestry and ASD case status.

This dissertation contributes to our understanding of the role of MIA exposure during pregnancy in ASD risk, biological changes identified in early childhood associated with prenatal exposure to MIA, and suitable methods for conducting EMR-based epidemiological research of ASD.

Advisor:	Christine Ladd-Acosta, PhD
Thesis Advisory Committee:	Terri Beaty, PhD, MA
	Li-Ching Lee, PhD, ScM
Readers:	M. Daniele Fallin, PhD
	Alden Gross, PhD
	Kristin Voegtline, PhD
Alternates:	Corinne Joshu, PhD
	Xiaobin Wang, MD, ScD

Acknowledgements

Thank you to my husband, Benjamin O'Neil. Even though this dissertation marks the end of my PhD, I am still just at the start of my training. There will be many more late nights and irregular hours; there will be hardships but also triumphs. I want to thank you for all the support you have given me, and all the support I know you will give me in the future, especially in the form of coffee.

Thank you to my advisor, Christine Ladd-Acosta, who has advocated for me at every step of this process. I am so grateful that you took a chance on me to be your first grad student. I am very glad that we will continue to work together and that this dissertation does represent the end of our relationship.

Thank you to the faculty who have served on my departmental and preliminary oral exams; thank you to the faculty and students who comprise the F(allin)L(add-Acosta)B(enke)V(olk) research group.

In particular, thank you to Shan Andrews. Though you may not know it, you basically taught me how to code. I would have floundered without your guidance.

Thank you to Hans Bjornsson. Working with the Kabuki Syndrome data marked a huge turning point in my skills as a programmer, statistician, and epigenetic epidemiologist; I would not be where I am today without that project and your mentorship.

Thank you to Bob Siliciano, Sharon Welling, Andrea Cox, Bern Harper, and Martha Buntin; the support of the MD-PhD program has sustained me for these past 7

years. I am very glad that I chose to come to Hopkins. I can only hope that I will be a credit to this program.

Thank you to Sophie Lanzkron. With your mentorship and guidance, I think I have a chance at being an excellent clinician and researcher; I hope to be as skilled, compassionate, dedicated and questioning as you. Your patients are lucky to have you.

Thank you to Emily and Diego Socolinsky, and the wonderful community of powerlifters I have somehow stumbled into; you have kept me grounded when I could have disappeared into this work.

Thank you to Dmitri, Spike, and Watson, on whom I can always count for comfort and respite and hijinks. Because my work is computational, I spent a lot of time on the couch with a dog asleep on my legs and a cat on my chest. I am very lucky.

Thank you to everyone that has become part of my family and my history: the greater O'Neil-Waldman clan; friends I have now known half my life—Krystle, Jen, Eric, Jeremy, Alex; friends who have known me as an adult—Andrew, Greg, Liana, Allie, Haley, Sarah, Courtney, and the entire matriarchy. My heart is full because of you.

"Social justice is the moral foundation, the heart, of public health." — Ruth Faden

Table of contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Autism Spectrum Disorder

1.1.1 Definition

Autism spectrum disorder (ASD) is characterized by deficits in social interaction and communication (both verbal and nonverbal), and repetitive behavior or stereotypical interests. ASD is increasingly common, with a prevalence in the United States of 1 in 68 children (1 in 42 boys and 1 in 189 girls) as of 2012 (Christensen, Baio et al. 2016). ASD is by definition a heterogeneous disease, with communication deficits and repetitive behaviors existing on a spectrum; the causes of ASD may be as diverse as its different presentations, with evidence for strong genetic and environmental contributions to disease risk (Hallmayer, Cleveland et al. 2011, Persico and Napolioni 2013, Sandin, Lichtenstein et al. 2014).

1.1.2 Etiology: genetics

The strong familial risk of ASD was first established through twin studies, comparing the autism concordance rate of dizygotic twins to monozygotic twins (Folstein and Rutter 1977, Steffenburg, Gillberg et al. 1989, Bailey, Le Couteur et al. 1995, Papadakis, Baltzis et al. 2011). Additionally, a high sibling recurrence risk has been demonstrated in families who already have a child diagnosed with ASD; the recurrence risk exists for a confirmed diagnosis of ASD, but also extends to a broader quantitative ASD phenotype. The sibling recurrence risk for a diagnosed ASD ranges from 10-20%,

but inheritance of quantitative traits related to ASD (primarily related to language or social communication) may extend to another 20% of siblings (Constantino, Zhang et al. 2010, Ozonoff, Young et al. 2011).

The strong genetic risk for ASD may be due to inherited common variants, inherited rare variants, or *de novo* rare variants. In some cases, *de novo* genetic mutations, particularly those that disrupt a protein-coding gene, or *de novo* copy number variants appear to explain the genetic risk for an individual (Sebat, Lakshmi et al. 2007, Glessner, Wang et al. 2009, Sanders, Ercan-Sencicek et al. 2011, Iossifov, Ronemus et al. 2012, Neale, Kou et al. 2012, Sanders, Murtha et al. 2012). There may be particular sites across the genome, or "hot spots," where rare *de novo* variants related to ASD risk independently arise more frequently than would be expected (Weiss, Shen et al. 2008). While in a few cases a specific genetic cause can be identified, ASD generally appears to be a polygenic disorder (Neale, Kou et al. 2012, O'Roak, Vives et al. 2012).

Although individually rare variants, particularly *de novo* variants, may be strongly associated with ASD, the additive effects of common variation probably explain much of the risk across a population, even though each common variant confers little risk by itself (Wang, Zhang et al. 2009, Anney, Klei et al. 2012, Klei, Sanders et al. 2012, Gaugler, Klei et al. 2014, Robinson, St Pourcain et al. 2016). One recent estimate suggests that the narrow-sense heritability (the ratio of additive genetic variance to the overall phenotypic variance) of ASD is approximately 50%, while the variance in liability due to rare, *de novo* variation is only 2.6% (Gaugler, Klei et al. 2014). Our understanding of the contribution of common variation to genetic ASD risk will improve as large

collaborations, such as that between the Psychiatric Genomics Consortium autism group (PGC-ASD) and the Danish iPsych project, increase our power to detect the common variants contributing to ASD risk (Robinson, St Pourcain et al. 2016).

While early twin studies estimated the genetic heritability of ASD to be as high as 90%, work that has taken into account shared and non-shared environmental factors suggests that the variance in ASD liability is substantially affected by both environment and genetics (Taniai, Nishiyama et al. 2008, Hallmayer, Cleveland et al. 2011, Sandin, Lichtenstein et al. 2014). Environmental exposures may explain more than 50% of the individual variance in disease risk, and might be particularly important in settings with a low prevalence rate, such as in sporadic or simplex families with only one ASD case (Tick, Bolton et al. 2016). Indeed, even early twin studies noted significant differences in the rate of obstetric complications between ASD affected and unaffected twins; the incidence of complications could be related to abnormal fetal development and placental dysfunction, or may represent an early environmental exposure to perinatal stress (Steffenburg, Gillberg et al. 1989, Bailey, Le Couteur et al. 1995).

1.1.3 Etiology: environment

Several demographic factors are strongly associated with ASD risk. Males are more commonly affected than females, at a ratio of approximately 4:1, though females with ASD are more likely to be severely affected and intellectually disabled (Werling and Geschwind 2013, Christensen, Baio et al. 2016). Ethnicity or ancestry may be related to ASD risk, with the highest prevalence observed among white rather than black or

Hispanic Americans as of 2012 (Christensen, Baio et al. 2016). In the US, foreign born mothers are also more likely to have children with ASD, as are more educated mothers (Guinchat, Thorsen et al. 2012, Dickerson, Rahbar et al. 2016). However, it is possible that the association with both race and education in the United States is related to socioeconomic status and services availability; social factors can influence the probability of ASD diagnosis (Mazumdar, Winter et al. 2013, Dickerson, Rahbar et al. 2016).

Increasing parental age (both maternal and paternal) is significantly associated with increased risk of ASD, as is being a young mother (Sandin, Schendel et al. 2016). There are several other pre- or peri-conception exposures that might be related to ASD risk. Studies have found that a short (<12 months between births) or long (≥72 months between births) inter-pregnancy interval is significantly associated with increased ASD risk in the child (Zerbo, Yoshida et al. 2015, Conde-Agudelo, Rosas-Bermudez et al. 2016), while the use of folic acid supplements prior to conception and early in pregnancy is protective (Smith, Strutton et al. 1990, Schmidt, Tancredi et al. 2012, Gao, Sheng et al. 2016). Maternal obesity and diabetes prior to pregnancy are also significantly associated with increased risk for ASD in the child (Li, Fallin et al. 2016).

A number of exposures during pregnancy have been implicated as potential risk factors for ASD later developing in the exposed fetus. These include dietary factors, such as folic acid supplementation as above or maternal vitamin D deficiency, which increases the risk of ASD with co-occurring intellectual disability (Magnusson, Lundberg et al. 2016). Toxicant exposures during pregnancy that have been implicated in ASD risk include air pollution (Volk, Lurmann et al. 2013, Flores-Pajot, Ofner et al. 2016, Lam,

Sutton et al. 2016) and pesticides (Roberts, English et al. 2007, Shelton, Geraghty et al. 2014, Lyall, Croen et al. 2017), while further work is needed to assess the risk associated with endocrine disrupting chemical exposure during pregnancy (Schmidt, Lyall et al. 2014). The use of medications during pregnancy, such as acetaminophen (Avella-Garcia, Julvez et al. 2016, Liew, Ritz et al. 2016) and SSRIs (Croen, Grether et al. 2011, Rai, Lee et al. 2013, Harrington, Lee et al. 2014) has also been linked to ASD risk, though this research is complicated by the possibility of confounding by indication when assessing risk associated with medication use.

Pregnancy complications such as preeclampsia are also associated with increased ASD risk (Getahun, Fassett et al. 2017). Neonatal ASD risk factors include preterm birth or being born small for gestational age, planned cesarean section, low Apgar scores, or hyperbilirubinemia (neonatal jaundice) (Guinchat, Thorsen et al. 2012, Logan, Dammann et al. 2017). In particular, the risk for ASD increases as a child is born at a lower gestational age (Joseph, Korzeniewski et al. 2017). Respiratory stress and hypoxia at birth have also been identified as ASD risk factors (Froehlich-Santino, Londono Tobon et al. 2014).

There may also be postnatal factors related to ASD risk, though fewer significant associations have been found compared to preconception or prenatal exposures. Early life exposure to air pollution has been implicated in ASD risk (Volk, Lurmann et al. 2013, Flores-Pajot, Ofner et al. 2016). Respiratory infections may be increased in children with ASD, though this was not assessed prospectively (Hadjkacem, Ayadi et al. 2016); it is, however, consistent with animal models that have shown early postnatal inflammation to

be associated with ASD-like behavior and brain overgrowth (Pang, Dai et al. 2016). While childhood vaccinations were once postulated to be an early life exposure that increased ASD risk, extensive research has demonstrated that this is not the case (Peterson and Barbel 2013, American Academy of Pediatrics 2017).

1.1.4 Prenatal exposures

Many of the identified environmental risk factors for ASD appear to act prior to or during pregnancy. It is plausible that exposures during pregnancy may alter neurodevelopment and increase disease risk (Rodier, Ingram et al. 1996, Rice and Barone 2000, Schlotz and Phillips 2009, Marques, O'Connor et al. 2013, Lyall, Schmidt et al. 2014), in part because children who will go on to be diagnosed with ASD may show differences in their behavioral and physical development very early in life. Children who will go on to be diagnosed with ASD may start to have an excessive increase in head circumference—evidence of brain overgrowth—as early as 1-2 months of age (Courchesne, Carper et al. 2003). Brain imaging directly shows this brain volume overgrowth, while a hyperexpansion of the cortical surface area can also be seen between 6 and 12 months of age (Hazlett, Gu et al. 2017). Children with ASD may show differences in fine motor and grasping skills as early as six months old, as well as head lag, an indicator of poor postural control (Libertus, Sheperd et al. 2014, Flanagan, Landa et al. 2012). There may be differences in gene expression in young children who will later be diagnosed with ASD (Glatt, Tsuang et al. 2012, Pramparo, Pierce et al. 2015). Primary health care professionals can be trained to accurately recognize potential signs of ASD,

related to social attention and communication, as early as 8 months of age (Barbaro and Dissanayake 2010).

In summary, current evidence indicates that ASD is a developmental disease that begins very early in life, and likely during gestation, as a result of genetic liability, environmental exposures, and the interaction between the two (Bakulski, Singer et al. 2014). Notably, because the prenatal period is a time where environmental exposures are most likely to influence later ASD development (Barouki, Gluckman et al. 2012, Stoner, Chow et al. 2014), this immediately suggests a population where timely interventions on environmental risk factors may act to decrease ASD incidence.

As described above, there are many potential prenatal exposures that could increase risk of autism in offspring (Gardener, Spiegelman et al. 2009). However, there is one significant risk factor that we have yet to mention: currently, a body of work from animal models and human epidemiological studies is coalescing to provide strong evidence that prenatal exposure to maternal immune activation (MIA) increases the risk of many different types of neurodevelopmental disorders (NDD), including autism.

1.2 Maternal Immune Activation

Maternal Immune Activation (MIA) is a construct representing increased maternal immune activity during pregnancy, as the result of an infection, inflammatory symptoms such as fever, or autoimmune activity. In particular, this dissertation focuses on MIA from

infection and fever exposure during gestation, primarily because they are common exposures (Collier, Rasmussen et al. 2009), likely with a high population attributable risk for ASD.

1.2.1 Congenital infections

One line of evidence is based on case reports of neurodevelopmental disease developing after congenital infections; autism was noted in some patients with congenital rubella (Chess 1971, Chess 1977, Chess, Fernandez et al. 1978, Chauhan, Sen et al. 2016), cytomegalovirus (Yamashita, Fujimoto et al. 2003, Garofoli, Lombardi et al. 2017), and herpes simplex virus (Ghaziuddin, Tsai et al. 1992). These early reports of increased prevalence of ASD in children with congenital infections led to further examination of the consequences of *in utero* infection exposure.

1.2.2 Analogy to schizophrenia

There are a number of ecological studies showing association between schizophrenia and winter/spring birth, when pregnancy would overlap with flu season; this potential association has been validated in studies that more specifically looked at serologically confirmed influenza during pregnancy and schizophrenia development in the offspring (Brown, Begg et al. 2004). Several different birth cohort studies have demonstrated associations between schizophrenia and maternal infections with *Toxoplasma gondii* (Brown, Schaefer et al. 2005, Mortensen, Norgaard-Pedersen et al. 2007) and herpes simplex virus type 2 (Buka, Tsuang et al. 2001a, Buka, Cannon et al.

2008). Elevated levels of maternal inflammatory mediators during pregnancy such as interleukin-8 (IL-8, renamed CXCL8) and tumor necrosis factor alpha (TNF-alpha) and the acute phase reactant C-reactive protein (CRP) are also associated with an increased risk of the exposed child developing schizophrenia (Buka, Tsuang et al. 2001b, Brown, Hooton et al. 2004, Canetta, Sourander et al. 2014). This work with maternal immune activation and schizophrenia serves as a proof of principle: *in utero* inflammation can alter neurodevelopment in ways that may not become clinically apparent until years after birth. Given the overlap of genetic and environmental risk factors for schizophrenia and autism (Meyer, Feldon et al. 2011), we are interested in the possibility that some of these same processes are involved in autism development. Evidence to support this hypothesis is accumulating from animal and human studies, as summarized below.

1.2.3 Animal models

Animal models of maternal immune activation generally rely on either direct injection of infectious antigens or induction of a generalized inflammatory response, and show behavioral, hormonal, neuropathologic, and transcriptional abnormalities in exposed offspring in species from mice to monkeys (Iwata, Matsuzaki et al. 2010, Harvey and Boksa 2012). Behavioral abnormalities noted in mice after exposure to prenatal inflammation include decreased vocalizations when separated from their mother or in social interactions as adults; decreased sociability; and increased repetitive behaviors in marble-burying and self-grooming tests (Malkova, Yu et al. 2012, Schwartzer, Careaga et al. 2013). Behavioral abnormalities have also been seen in rhesus monkeys who were

exposed to MIA by experimentally triggered systemic inflammation. MIA exposed monkeys exhibited increased repetitive behaviors, decreased social affiliative behaviors, and abnormal gaze patterns in response to faces (Bauman, Iosif et al. 2014, Machado, Whitaker et al. 2015). Abnormalities in immune function, hormonal differences, and altered neurotransmitter levels have also been noted in mice exposed to MIA (Hsiao, McBride et al. 2012, Miller, Zhu et al. 2013, Ohkawara, Katsuyama et al. 2015). Impaired neuronal growth has been observed after exposure to MIA (Straley, Togher et al. 2014). Cerebellar abnormalities or mild brain overgrowth after MIA exposure, closely mirroring what is seen in humans with ASD, have been detected as well (Shi, Smith et al. 2009, Le Belle, Sperry et al. 2014). It appears that elevated levels of interleukin-6, a pro-inflammatory cytokine, are required for the pathologic effects of MIA to manifest in exposed offspring (Smith, Li et al. 2007); this work has been extended to demonstrate that IL-6 works upstream of IL-17a, an effector cytokine that is released from specific T helper cells (T_H17 cells) and is responsible for the abnormal cortical phenotype in MIA-exposed mouse pups (Choi, Yim et al. 2016). The ability to induce MIA in experimental models independent of particular infectious antigens demonstrates that the negative effects of prenatal MIA are not directly caused by the infectious organism itself, but by the maternal immune response and systemic inflammation. In general, data derived from animal models consistently points to deranged neurodevelopment after exposure to MIA, in ways that are causally related to behavioral and pathologic changes seen in schizophrenia and autism.

1.2.4 Epidemiologic studies

Human observational studies have also demonstrated a relationship between inflammatory biomarkers during pregnancy and autism risk in the offspring. Amniotic fluid obtained during pregnancy was shown to differ in chemokine and cytokine levels between children who went on to develop ASD and those who did not; particularly striking is that even when controlling for maternal autoimmunity and history of infections during pregnancy, fetuses that went on to develop ASD after birth were exposed to significantly higher levels of TNF-alpha and TNF-beta (Abdallah, Larsen et al. 2012, Abdallah, Larsen et al. 2013). In a case-control study with access to archived maternal serum samples, mothers of children with ASD had significant elevations in mid-pregnancy levels of interferon-gamma, IL-4, and IL-5 (Goines, Croen et al. 2011). A recent study found that mothers whose children developed ASD with intellectual disability had higher mid-gestational levels of granulocyte macrophage colony-stimulating factor, interferon-gamma, interleukin-1alpha, and IL-6, compared to the mid-gestational levels of mothers whose children were typically developing (Jones, Croen et al. 2017).

C-reactive protein (CRP) is an acute phase reactant that is often used as a clinical marker of significant systemic inflammation; a study of 677 ASD cases and 677 matched controls derived from the Finnish Maternity Cohort found a significant association between CRP during pregnancy and risk of the offspring developing ASD. Mothers in the highest quintile of CRP values had an OR of 1.43 [1.02-2.01] of their child developing

ASD as compared to mothers in the lowest quintile of CRP values (Brown, Sourander et al. 2014).

A study of all children born in Denmark from 1980 to 2005 (n = 1,612,342) took advantage of the ability to link the Danish Medical Birth Register, Danish Psychiatric Central Register, and Danish National Hospital Register to identify children whose mother was hospitalized with an infection while pregnant with them. A mother's admission to a hospital with a viral infection in her first trimester or a bacterial infection in her second trimester was significantly associated with ASD risk in the offspring (adjusted hazard ratio of 2.98 [95% CI 1.29-7.15] for viral infection in first trimester, and adjusted hazard ratio of 1.42 [1.80-1.87] for bacterial infection in the second trimester) (Atladottir, Thorsen et al. 2010). A study from the Danish National Birth Cohort (n = 101,033) had self-reported infection, fever, and antibiotic use data, prospectively collected at 17 and 32 weeks gestation. They reported a significantly increased risk of ASD in children exposed to a prolonged period of maternal fever *in utero* (adjusted hazard ratio, 3.2 [1.8-5.6]). Similar to the 2010 study, they also found that maternal self-report of influenza (a viral infection) during the first trimester was significantly associated with early-onset autism in the children (Atladottir, Henriksen et al. 2012). However, an HMO-based case-control study found no association between maternal influenza—defined by a medical record diagnosis of influenza or a positive influenza laboratory result—and risk of ASD (Zerbo, Iosif et al. 2013). A large autism case-control study in California with rigorously validated ASD outcomes confirmed the association between self-reported maternal fever during pregnancy and autism (OR 2.12 [1.17-3.84])

and other developmental disabilities (OR 2.5 [1.20-5.20]). This association was attenuated in mothers who took anti-pyretic medications to control their fever, but remained elevated in mothers who did not (OR 2.55 [1.30-4.99]) (Zerbo, Iosif et al. 2013). A nested case-control study using the Kaiser Permanente Northern California database found an association between hospitalization with an infection during pregnancy and increased ASD risk (adjusted OR 1.48 [1.07-2.04]), with the risk elevated to almost 60% above baseline for those women hospitalized with a bacterial infection. This study also noted that having multiple hospitalizations for infections was associated with increased ASD risk (adjusted OR 1.36 [1.05-1.78]) (Zerbo, Qian et al. 2013). A large Swedish registry study (n = 2,371,403) demonstrated an association between maternal inpatient diagnosis of infection any time during pregnancy and ASD in the offspring (OR 1.37 [1.28-1.47]). This analysis was robust to an assumption of residual confounding and persisted for all infection categories regardless of organism, infection site, or trimester of infection (Lee, Magnusson et al. 2014). A study in Taiwan using incident ASD cases and matched controls from a health insurance database found that two or more maternal outpatient visits for a genital infection (aOR 1.34 [1.12-1.60]) or bacterial infection (aOR 1.25 [1.06-1.43]) during the third trimester of pregnancy was significantly associated with increased ASD risk (Fang, Wang et al. 2015). In a Norwegian birth cohort, high levels of IgG antibodies against herpes simplex virus-2 during midpregnancy was significantly associated with an increased risk of ASD, but only in male offspring (aOR 2.07 [1.05-4.06]) (Mahic, Mjaaland et al. 2017). Finally, a recent meta-analysis found that any maternal infection during pregnancy was modestly associated with an increased

risk of ASD, and that this may be a stronger association in those hospitalized for an infection during pregnancy (Jiang, Xu et al. 2016).

There is also evidence that prenatal infections appear to increase autism severity in individuals who are otherwise genetically susceptible by virtue of pathologic copy number variations (Mazina, Gerdts et al. 2015). This is in concordance with earlier evidence that familial liability and exposure to prenatal infection interact to increase schizophrenia risk (Clarke, Tanskanen et al. 2009), suggesting that there may be classes of individuals who are particularly susceptible to the effects of prenatal MIA.

1.3 Focus of this dissertation

With strong evidence for an association between MIA and neurodevelopmental disabilities in general and ASD specifically, further work to understand the biological consequences of prenatal exposure to MIA in humans is required. Infection and fever are both very common in pregnancy, with a prevalence of self-reported infection of 63.6% in one US cohort, and a prevalence of fever of 20.5% (Collier, Rasmussen et al. 2009). Understanding the consequences of prenatal infection and fever and their role in neurodevelopmental deficits is thus of clear public health importance. We are particularly interested in the role that epigenetics might play in either mediating or marking gestational infection exposure.

Epigenetics is the study of DNA regulation that is mitotically heritable, or passed from one cell to its daughter cell; one type of epigenetic control is DNA methylation, where a methyl group added to a specific DNA nucleotide can influence the expression of a gene either locally or at-a-distance. These methyl marks on DNA can differ over time, across cell types, and in response to the environment. DNA methylation can be assayed across the genome with an affordable platform, the Illumina Infinium HumanMethylation450 BeadChip methylation array ("450k platform"), which queries the degree of methylation at 485,512 genetic loci (Bibikova, Barnes et al. 2011). Using this platform allows us to test for association of methylation at any measured locus with a phenotype of interest, in our case history of prenatal exposure to MIA (Chadwick, Sawa et al. 2015). Previous work using this technology has shown that prenatal insults can epigenetically alter offspring. For example, prenatal exposure to tobacco smoke changes a newborn's DNA methylation in a set of locations across the genome, which forms an epigenetic signature reflecting prior exposure (Breton, Byun et al. 2009, Joubert, Haberg et al. 2012). This epigenetic signature of prenatal smoking exposure actually persists through the first few years of childhood, and can be detected in children 3-5 years of age (Ladd-Acosta, Shu et al. 2016). Prenatal exposure to infection may also lead to observable epigenetic changes in exposed children.

There are already tantalizing clues that epigenetics will be able to explain aspects of autism etiology. Epigenetic differences have been observed in the brains of individuals with ASD compared to normal controls (Schanen 2006, Shulha, Cheung et al. 2012, Ladd-Acosta, Hansen et al. 2014, Nardone, Sams et al. 2014, Loke, Hannan et al. 2015).

These differences have been detected and replicated with independent technologies and across relevant neuroanatomic regions, including those associated with language and movement. Additionally, evidence demonstrating epigenetic differences in the brains of adolescent mice exposed to inflammation *in utero* has emerged (Basil, Li et al. 2014). Epigenetic differences were also present in the medial prefrontal cortex of adult mice who had been exposed to MIA during gestation, in both a candidate region and genome-wide approach (Labouesse, Dong et al. 2015, Richetto, Massart et al. 2017). Similar changes might arise in the human brain, and lead to changes that are detectable in blood.

In summary, there is extensive evidence indicating that prenatal exposure to maternal immune activation (MIA), as a consequence of infection or fever, can play a role in the development of Autism Spectrum Disorder (ASD). In this dissertation, we estimate the association between MIA and ASD in a prospective birth cohort (**Chapter 2**) and a case-control study (**Chapter 4**), both in the United States. We also explore ways to improve ASD case identification in electronic medical records to assist future epidemiological studies of ASD risk factors (**Chapter 3**). We then conduct an Epigenome Wide Association Study (EWAS) to detect differential methylation in the peripheral blood of 2-5 year old children who were prenatally exposed to maternal infection (**Chapter 4**). Taken together, the work presented here extends research on the relationship between ASD risk and exposure to maternal immune activation.

Chapter 2: Prenatal exposure to fever is associated with Autism Spectrum Disorder in the Boston Birth Cohort

This chapter describes work currently under revision at Autism Research, with contributions from co-authors Christine Ladd-Acosta, Mengying Li, Deanna Caruso, Xiumei Hong, Jamie Kaczaniuk, Elizabeth A. Stuart, M. Daniele Fallin, and Xiaobin Wang.

2.1 Abstract

Autism Spectrum Disorder (ASD) is phenotypically and etiologically heterogeneous, with evidence for genetic and environmental contributions to disease risk. Research has focused on the prenatal period as a time when environmental exposures are likely to influence risk for ASD. Epidemiological studies have shown significant associations between prenatal exposure to maternal immune activation (MIA), caused by infections and fever, and ASD. However, due to differences in study design and exposure measurements no consistent patterns have emerged revealing specific times or type of MIA exposure that are most important to ASD risk. As well, no prior studies have examined prenatal MIA exposure and ASD risk in a predominantly under-represented minority population. To overcome these limitations we estimated the association between prenatal exposure to fever and maternal infections and ASD in a prospective birth cohort

of an understudied urban minority population in the United States. No association was found between prenatal exposure to genitourinary infections or influenza and the risk of ASD in a nested sample of 116 ASD cases and 988 typically developing controls in crude or adjusted analyses. Prenatal exposure to fever was associated with increased ASD risk (aOR = 2.02 [1.04 – 3.92]) after adjustment for educational attainment, marital status, race, child sex, maternal age, birth year, gestational age and maternal smoking. This effect may be specific to fever during the third trimester (aOR 2.70 [1.00 – 7.29]). Our findings provide a focus for future research efforts and ASD prevention strategies across diverse populations.

2.2 Introduction

Autism spectrum disorder (ASD) is characterized by deficits in social interaction or communication and repetitive behavior or stereotypical interests. ASD is increasingly common, with a prevalence of 1 in 68 children (1 in 42 boys and 1 in 189 girls) as of 2012 (Christensen et al. 2016). ASD is phenotypically and etiologically heterogeneous, with evidence for both genetic and environmental contributions to disease risk (Hallmayer et al. 2011, Persico et al. 2013, Sandin et al. 2014). Converging evidence points to the prenatal period as a time when environmental exposures are most likely to influence ASD risk (Rodier et al. 1996, Rice et al. 2000, Schlotz et al. 2009, Barouki et al. 2012, Marques et al. 2013, Lyall et al. 2014, Stoner et al. 2014). Identification of

modifiable ASD risk factors can lead to preventative intervention strategies that may reduce overall ASD burden. Studies to examine a broad range of environmental risk factors for ASD during this critical time window are now emerging.

There is a growing body of evidence suggesting prenatal exposure to maternal immune activation (MIA) and/or systemic inflammation increases the risk of many different types of neurodevelopmental disorders (NDD), including autism. Animal models of MIA have shown behavioral, hormonal, and neuropathologic differences among prenatally exposed offspring relative to their unexposed counterparts (Malkova et al. 2012, Miller et al. 2013, Schwartzer et al. 2013, Bauman et al. 2014, Machado et al. 2015). In addition, MIA-associated differences in immune function (Hsiao et al. 2012), hormone and neurotransmitter levels (Miller et al. 2013, Ohkawara et al. 2015), neuronal and whole brain growth (Shi et al. 2009, Le Belle et al. 2014, Straley et al. 2014), as well as microglial neurodevelopmental regulatory patterns (Miller et al. 2013, Matcovitch-Natan et al. 2016) have been found in animal models and are consistent with observations in humans with ASD.

Human studies have identified associations between prenatal exposure to maternal infection and ASD risk. Two European registry-based population studies and one US HMO-based case-control study have identified associations between maternal infection, bacterial or viral, during pregnancy and increased ASD risk in her offspring (Atladottir et al. 2010, Lee et al. 2015), with the highest elevated risk among women with multiple hospitalizations for infections or those with bacterial infections (Zerbo et al. 2015). One case-control study in Taiwan using medical record data found an elevated risk for ASD

after genital or bacterial infections (Fang et al. 2015). In addition, a study that assessed maternal exposure to infection by self-report, rather than medical records, showed potential risk effects for influenza (Atladottir et al. 2012), although a recent HMO-based cohort analysis did not observe such an association (Zerbo et al. 2017). A meta-analysis of 15 studies found an increase in ASD risk after any type of maternal infection (Jiang et al. 2016).

Fewer studies have examined the potential impact of fever specifically, rather than infection broadly, on ASD risk. One study found that prolonged febrile episodes were associated with increased risk (Atladottir et al. 2012). A retrospective case-control study based on maternal self-report showed association between fever during pregnancy and increased ASD risk (Zerbo et al. 2013). That study further showed that risk was attenuated in mothers who took anti-pyretic medications to control their fever, but remained elevated in mothers who did not (Zerbo et al. 2013). Fever exposure has also been shown to adversely influence developing fetal health more generally (Dreier et al. 2014).

Effect sizes for associations between MIA exposure and ASD have been relatively modest; however, infection during pregnancy is common and thus can have a major impact on disease burden. The prevalence of self-reported infection during pregnancy in the US has been reported to be as high as 63.6%, and 20.5% for fever (Collier et al. 2009). While studies to date suggest prenatal MIA as an important environmental risk factor for ASD, there remain several limitations with respect to study design, generalizability, time resolution, ability to examine specific infectious agents or organ

systems, and lack of inclusion of minority groups. While we were not able to address all of the limitations of the literature, here we overcome concerns associated with non-diverse samples and retrospective exposure assessment by performing a prospective analysis of prenatal exposure to MIA and ASD risk in an under-represented minority birth cohort in the US.

2.3 Methods

2.3.1 Boston Birth Cohort (BBC) Study Description

Mother-child pairs were from the Boston Birth Cohort (BBC), a prospective birth cohort with pregnancy exposure, early life factors, and phenotypic data available for over 8,000 mother-child dyads recruited at the Boston Medical Center (Wang et al. 2002, Wang et al. 2014). The BBC enrolls predominantly urban, low-income minority mothers and their children; the representative subsample examined in this study is approximately 38% black/African American, 22% Hispanic, 19% Haitian, and 8.5% white (see Table 2.1). The majority of BBC mothers are receiving health care through public assistance based insurance programs, e.g. Medicaid or MassHealth. The Boston Birth Cohort was established as the Molecular Epidemiology of Preterm Delivery in 1998 with the recruitment of women delivering at the Boston Medical Center (BMC), with oversampling for preterm birth. Women with a singleton live birth at the BMC are eligible for recruitment, with exclusions for IVF, multiple gestations, chromosomal

abnormalities, major birth defects, and preterm deliveries due to maternal trauma. As described in detail in Wang et al. (2002), participants are contacted 24-72 hours after birth to obtain consent and initiate study enrollment. Starting in 2002, child development was followed through electronic medical records for the subset of BBC children who received pediatric care at the Boston Medical Center (n=2992). At the time of recruitment into the BBC, each mother-child dyad is given a unique study ID that is linked to the electronic medical record identifiers for both the mother and child.

Table 2.1: Characteristics of the ASD case-control study sample from the Boston Birth Cohort (BBC)

	Neurotypical Controls (*n*=988)	ASD (*n*=116)	*P*
Gravidity, M (SD)	2.79 (1.79)	2.85 (1.93)	0.701
Parity, M (SD)	1.04 (1.20)	0.96 (1.22)	0.491
Maternal age[a], M (SD)	28.25 (6.55)	30.11 (6.24)	0.004*
Education, *n* (%)			0.281
Elementary school	41 (4.3)	4 (3.5)	
Secondary school	238 (24.7)	21 (18.6)	
High school/GED	333 (34.6)	40 (35.4)	
Some college	203 (21.1)	33 (29.2)	
College degree and above	148 (15.4)	15 (13.3)	
Marital status, *n* (%)			0.570
Married	327 (34.0)	42 (37.5)	
Not Married	661 (66.0)	74 (62.5)	
Race or Ethnicity[b], *n* (%)			0.472
Black	615 (62.2)	73 (62.9)	
White	83 (8.4)	6 (5.2)	
Hispanic	214 (21.7)	31 (26.7)	
Asian	20 (2.0)	2 (1.7)	
Other	56 (5.7)	4 (3.4)	
Maternal smoking[c], *n* (%)			0.179
Never	817 (85.1)	88 (78.6)	
Some	48 (5.0)	9 (8.0)	
Continuous	95 (9.9)	15 (13.4)	
Child sex, *n* (%)			<0.001*
Female	585 (59.2)	31 (26.7)	

Male	403 (40.8)	85 (73.3)	
Gestational age, mean (SD)[d]	38.4 (2.7)	36.5 (4.6)	<0.001*
Child Year of birth, M (SD)	2006.0 (3.8)	2006.4 (3.5)	0.279
Birth weight, *n* (%)			0.072
>2500 grams	763 (78.9)	81 (71.1)	
<2500 grams	204 (21.1)	33 (28.9)	

ASD, autism spectrum disorder; BBC, Boston Birth Cohort; M, mean; SD, standard deviation

[a] Maternal age at time of delivery

[b] Black includes self reported Black, African American, Haitian, Cape Verdean, and Caribbean race and ethnicities. Asian includes Asian and Pacific Islander races. The Other category includes individuals with a mixed or other racial background.

[c] Never smokers were defined as mothers with no history of smoking 6 months prior to conception or during pregnancy; some smoking includes mothers that smoked at some point in the window of 6 months prior to conception and through delivery but did not smoke throughout that window; continuous is defined as mothers that smoked starting 6 months prior to and throughout pregnancy.

[d] Defined by sonogram

* Denotes statistically significant

2.3.2 Analytic Sample & Outcome Classification

We performed a case-control analysis of children with Autism Spectrum Disorder (ASD) and with neurotypical development. We used electronic medical record ICD-9-CM diagnosis codes for pediatric inpatient, outpatient, and emergency room visits to the Boston Medical Center, between 1 October 2003 and 30 September 2015 (the last date before transition from ICD-9-CM to ICD-10-CM), to define ASD cases and neurotypical controls. Specifically, individuals were classified as an ASD case if their medical records contained any of the following ICD-9-CM codes, at least once: 299.00, 299.01, 299.80, 299.81, 299.90, or 299.91. We classified individuals as neurotypical controls if they were

never diagnosed with any of the following conditions: ASD, attention deficit hyperactivity disorder (ADHD), intellectual disability (ID), developmental delay (DD), oppositional defiant disorder (ODD) or other "emotional disturbances of childhood," conduct disorder (CD), or congenital anomalies (based on ICD-9-CM codes; see Table 2.2).

Table 2.2: ICD-9-CM code based definitions for ASD cases and typically developing controls in the Boston Birth Cohort, 2003 – 2015

	ICD-9-CM codes	N
ASD case definition		120
Inclusion criteria (any one of these codes):	299.0, 299.01, 299.8, 299.81, 299.9, 299.91	
Neurotypical control definition		1033
Exclusion criteria (any one of these codes):		
ADHD	314.0 - 314.9	
Conduct Disorder	312.0 - 312.9	
Emotional disturbances of childhood or adolescence including Oppositional Defiant Disorder	313.0 - 313.9	
Developmental Delay	315.0 - 315.9	
Intellectual Disability	317 - 319	
Congenital Anomalies	740 - 759.9	

2.3.3 Exposure Definitions

Enrolled BBC mothers were interviewed 24-72 hours after delivery using a standardized postpartum questionnaire to gather information about her pregnancy (Wang et al. 2014). Data on prenatal exposure to influenza, fever (excluding intrapartum), and

genitourinary tract infections were obtained from self-report based on the questionnaire. In addition, history of an intrapartum fever (>38C) was abstracted from electronic medical record data by trained study personnel using a standardized form.

For each type of exposure examined, including prenatal genitourinary (GU) infections, prenatal influenza infection, maternal fever during pregnancy, and intrapartum maternal fever, we generated a dichotomous categorical variable representing "exposed" or "unexposed." Children whose mothers responded 'yes' to having any vaginal or genital tract or urinary tract infections during this pregnancy (including yeast infections), any fever during this pregnancy, and any flu during this pregnancy were defined as "exposed" for GU, fever, and flu variables, respectively (Table 2.3). Similarly, dichotomous trimester-specific variables were derived for flu and fever exposures using trimester-specific information among the subset of mothers that positively responded to having an exposure at any point during pregnancy. For intrapartum fever exposure, individuals were categorized as "exposed" if their mother had an intrapartum temperature > 38C, obtained via abstracted labor and delivery electronic medical records.

Table 2.3: Study questions about flu, fever, and infection exposure during pregnancy from the BMC Maternal at Postpartum questionnaire

Question	Answer
Did you have any flu during this pregnancy?	*Yes/no*
a. First trimester	*Yes/no*
b. Second trimester	*Yes/no*
c. Third trimester	*Yes/no*
Did you have any fever during this pregnancy?	*Yes/no*
a. First trimester	*Yes/no*
b. Second trimester	*Yes/no*
c. Third trimester	*Yes/no*
Did you have any vaginal or genital tract or urinary tract infections during pregnancy? (including yeast infections)	*Yes/no*
Examples:	
Chlamydia, Gonorrhea, Syphilis, Trichomonas, GBS, BV, Yeast, Herpes, HPV, Other GT, Unknown GTI, Urinary Tract	

2.3.4 Covariate Definitions

Covariates used for adjustment included characteristics of the mother (educational attainment, marital status, race, pregnancy smoking status, age at delivery) and child (sex, birth year, gestational age at birth according to ultrasound dating). Educational attainment, marital status, race, and smoking during pregnancy were all self-reported in the postpartum questionnaire. Child sex was self-reported and confirmed in the abstracted medical records. Maternal age at the time of delivery was self-reported and confirmed based on mother's date of birth recorded in her medical records. We defined gestational

age at birth using dating from the first available ultrasound in the medical records (Wang et al. 2014).

We defined mothers' educational attainment as a categorical variable (elementary school, secondary school, high school/GED, some college, or college/postgraduate degree) using the self-reported questionnaire data. Marital status was self-reported as married, single, divorced, separated, or widowed; this was dichotomized to "married" or "not married" for our analyses. Race was self-reported by checking one of the 9 following categories that best reflected the respondent's background: Black/African American; Asian; Pacific Islander; White; Haitian; Hispanic; Cape Verdean; Other; and Unknown. For our analyses, we then collapsed these responses into five categories to generate the race covariate: (1) Black, (2) White, (3) Hispanic, (4) Asian, and (5) all others. Black includes self-reported Black, African American, Haitian, Cape Verdean, and Caribbean race and ethnicities. White includes all individuals that reported white. Asian includes Asian and Pacific Islander races. The Other category includes all other backgrounds. Maternal smoking was a categorical covariate defined using self-reported data; never smokers were defined as mothers with no history of smoking 6 months prior to conception or during pregnancy. Mothers that smoked at any point in the 6 months prior to conception or at any point during their pregnancy were coded as "some smoking." Continuous smokers were defined as mothers that smoked 6 months prior to conception and throughout their pregnancy.

2.3.5 Statistical Analyses

All data cleaning and analysis was performed with R-3.1.3. Summary tables of characteristics of ASD cases and controls, as well as exposed and unexposed, were created with the R package *tableone* (https://CRAN.R-project.org/package=tableone). Descriptive statistics for categorical variables were obtained with the function chisq.test() with continuity correction, and the function oneway.test() for continuous variables with an assumption of equal variance.

ASD odds ratios for MIA exposures were estimated via unadjusted and adjusted binomial logistic regression using R-3.1.3. For each of the 4 exposures (prenatal fever, flu, GU infection, and intrapartum fever) we performed independent analyses. Our final model was adjusted for socioeconomic status as represented by mothers' educational attainment, marital status, and race, as well as for child sex, maternal age, child birth year and gestational age at birth, and maternal smoking during pregnancy. Beta coefficients from the logistic regression were transformed to obtain odds ratios for association with ASD outcome. A p-value < 0.05 was taken to be evidence for a statistically significant association. Forest plots were generated with the R package *metafor* (https://CRAN.R-project.org/package=metafor).

To allow a cleaner comparison of exposed and unexposed groups and to better take confounding into account, we also conducted a propensity score analysis for exposure to fever. We matched individuals who were exposed to maternal fever any time during their gestation with those who were not exposed to fever by their propensity for exposure to fever, conditional on maternal age, smoking status, race, and educational and

marital status, as well as child sex, gestational age at birth, and year of birth, using the R package *MatchIt* (Ho et al. 2007, Ho et al. 2011); exposed individuals were matched to their nearest unexposed, with a caliper of 0.08 and a case:control ratio of 1:8.

Probabilistic sensitivity analyses for the effect of exposure misclassification were performed using the R package *episensr* (https://CRAN.R-project.org/package=episensr) and the function probsens(), which implements the quantitative bias methods described in Lash et al. 2009. We explored various prior probability distributions of the sensitivity and specificity for fever, including uniform and logit-normal. We also used the functions misclassification() and boot.bias() to empirically derive, with bootstrapping, estimates of the uncertainty in the OR for different models of sensitivity and specificity, including differential and non-differential exposure misclassification.

2.4 Results

2.4.1 Sample Description

Electronic medical records for 121,457 inpatient, outpatient, and emergency room visits to the Boston Medical Center (BMC) were available for children enrolled in the Boston Birth Cohort follow-up study. After 2,436 visits contributed by siblings were removed from the dataset, 118,939 records from 2,992 index children remained; these visits occurred between 1 October 2003 and 30 September 2015. On average each child had 39.8 visits to the BMC, with a range of 1 to 463 visits.

120 ASD cases and 1033 neurotypical controls were identified in the dataset based on ICD-9-CM diagnoses, as described above. Mother-child pairs were constructed by linking maternal demographic data to child EMR data via a unique family-level study ID; 2 ASD cases and 22 controls were dropped at this stage. Then each mother-child pair was linked to questionnaire data on prenatal exposures via study ID. 23 controls and 2 cases were removed, leaving 988 neurotypical controls and 116 cases before listwise deletion.

There was only limited correlation between the exposure variables (Table 2.4). The highest correlation was seen between reported exposure to influenza and reported exposure to fever ($r = 0.31$).

At this stage, there was a low prevalence of missing data in the covariates (ranging from 0% for maternal age and child sex and birth year, to 2.1% for maternal race, and 4.3% for gestational age) and exposures of interest (1.4% missing for genitourinary infections, 2.6% for fever any time during pregnancy, 2.7% for flu any time during pregnancy, and 8.4% for intrapartum fever). We handled missingness by listwise deletion—removing a subject from further analysis if they were missing data—for each exposure and the set of covariates included in the final model (maternal education, marital status, race, age, and pregnancy smoking status; child sex, birth year, and gestational age at birth) (Table 2.5).

Table 2.4: Pearson's product-moment correlation between exposure variables

	GU	I	I, T1	I, T2	I, T3	IF	F	F, T1	F, T2	F, T3
Genito-urinary infections	1									
Influenza	0.004	1								
Trimester 1 (I, T1)	0.027	0.48	1							
Trimester 2 (I, T2)	0.022	0.60	0.092	1						
Trimester 3 (I, T3)	-0.038	0.65	0.14	0.095	1					
Intrapartum fever	0.037	0.013	0.015	0.0023	0.041	1				
Fever	0.034	0.31	0.18	0.17	0.13	-0.021	1			
Trimester 1 (F, T1)	0.029	0.13	0.32	-0.052	0.018	-0.022	0.58	1		
Trimester 2 (F, T2)	0.033	0.24	-0.016	0.39	-0.016	-0.045	0.57	0.029	1	
Trimester 3 (F, T3)	0.029	0.16	0.055	-0.052	0.24	0.031	0.58	0.058	-0.0014	1

Table 2.5: Sample sizes remaining after listwise deletion for each exposure model

exposure categories	neurotypical controls	ASD cases
GU infections	890	101
Influenza, any time during pregnancy	884	101
Influenza, trimester-specific	881	101
Fever, any time during pregnancy	884	101
Fever, trimester-specific	881	101
Intrapartum fever	843	95

Individuals in the final analytic sample and those removed by listwise deletion differed significantly by average year of birth. Specifically, for analysis of prenatal exposure to fever at any time during gestation, those retained in our analytic sample were born in 2006, on average, and those removed from the final analytic dataset were born in 2004, on average ($p = 0.001$). Otherwise, there were no significant differences between individuals retained for further analysis and those removed due to missing data (Table 2.6).

As expected, we observed significant associations between known autism spectrum disorder (ASD) risk factors including child male sex, increased maternal age, and lower gestational age in the Boston Birth Cohort (BBC) sample (Table 2.1). No significant ASD case-control differences were observed for maternal gravidity, parity, education, marital status, race/ethnicity, smoking status, or child birth weight or birth year (Table 2.1).

Table 2.6: Characteristics of mother-child pairs in the Boston Birth Cohort (BBC) who were missing data for prenatal exposure to fever at any time during pregnancy; 2.6% of the sample was missing data for prenatal fever exposure

	Percent Missing	No missingness (n=985)	Missing data (n=119)	*p*
ASD diagnosis, *n* (%)	0%			0.528
no		884 (89.7)	104 (87.4)	
yes		101 (10.3)	15 (12.6)	
Maternal age[a], M (SD)	0%	28.46 (6.54)	28.29 (6.59)	0.780
Child year of birth, M (SD)	0%	2006.18 (3.52)	2004.93 (5.48)	0.001*
Education, *n* (%)	2.5%			0.853
Elementary school		40 (4.1)	5 (5.5)	
Secondary school		240 (24.4)	19 (20.9)	
High school/GED		339 (34.4)	34 (37.4)	
Some college		215 (21.8)	21 (23.1)	
College degree and above		151 (15.3)	12 (13.2)	
Marital status, *n* (%)	2.6%			0.138
Married		640 (65.0)	66 (73.3)	
Not Married		345 (35.0)	24 (26.7)	
Race or Ethnicity[b], *n* (%)	2.1%			0.563
Black		605 (61.4)	60 (62.5)	
White		84 (8.5)	5 (5.2)	
Hispanic		219 (22.2)	26 (27.1)	
Asian		21 (2.1)	1 (1.0)	
Other		56 (5.7)	4 (4.2)	
Maternal smoking[c], *n* (%)	2.9%			0.788
Never		833 (84.6)	72 (82.8)	

Some		51 (5.2)	6 (6.9)	
Continuous		101 (10.3)	9 (10.3)	
Child sex, *n* (%)	0%			0.984
Female		549 (55.7)	67 (56.3)	
Male		436 (44.3)	52 (43.7)	
Gestational age, mean (SD)[d]	4.3%	38.22 (3.03)	37.97 (3.13)	0.504
Birth weight, *n* (%)	2.1%			
>2500 grams		215 (21.8)	22 (22.9)	0.907
<2500 grams		770 (78.2)	74 (77.1)	

ASD, autism spectrum disorder; BBC, Boston Birth Cohort; M, mean; SD, standard deviation

[a] Maternal age at time of delivery

[b] Black includes self reported Black, African American, Haitian, Cape Verdean, and Caribbean race and ethnicities. Asian includes Asian and Pacific Islander races. The Other category includes individuals with a mixed or other racial background.

[c] Never smokers were defined as mothers with no history of smoking 6 months prior to conception or during pregnancy; some smoking includes mothers that smoked at some point in the window of 6 months prior to conception and through delivery but did not smoke throughout that window; continuous is defined as mothers that smoked starting 6 months prior to and throughout pregnancy.

[d] Defined by sonogram

* Denotes statistically significant

2.4.2 Prenatal Exposure to Genitourinary Infection and ASD Risk

No association was found between self-reported maternal history of genitourinary (GU) infections at any time during pregnancy and risk of ASD development in the offspring in an unadjusted model (OR 0.83 [95% confidence interval 0.52 – 1.32]; Table 2.7). Similarly, no significant association was observed after adjusting for child sex, maternal age, child birth year, maternal smoking status, maternal education, marital status, maternal race, and gestational age (aOR 0.69 [0.42 – 1.14]; Figure 2.1 and Table 2.7). Our findings in the BBC sample are consistent with three prior studies (Atladottir 2010, Lee 2015, Fang 2015) that also examined prenatal exposure to GU infections and ASD risk (Figure 2.2).

Table 2.7: Odds ratios and 95% confidence intervals for the association between all tested exposures and ASD in the Boston Birth Cohort

exposure categories	OR [95% CI]
GU infections, unadjusted	0.83 [0.52 - 1.32]
GU infections, adjusted	0.69 [0.42 - 1.14]
Influenza, unadjusted	1.14 [0.69 - 1.89]
Influenza, adjusted	1.17 [0.67 - 2.05]
in first trimester, unadjusted	0.88 [0.34 - 2.27]
in first trimester, adjusted	0.93 [0.33 - 2.59]
in second trimester, unadjusted	1.69 [0.88 - 3.25]
in second trimester, adjusted	1.34 [0.65 - 2.77]
in third trimester, unadjusted	0.83 [0.39 - 1.76]
in third trimester, adjusted	0.95 [0.42 - 2.12]
Intrapartum fever, unadjusted	0.52 [0.16 - 1.69]
Intrapartum fever, adjusted	0.60 [0.18 - 2.02]
Any fever, unadjusted	1.80 [0.99 - 3.27]
Any fever, adjusted	**2.02 [1.04 - 3.92]**
in first trimester, unadjusted	1.21 [0.42 - 3.52]
in first trimester, adjusted	1.86 [0.61 - 5.73]
in second trimester, unadjusted	2.26 [0.90 - 5.66]
in second trimester, adjusted	1.62 [0.58 - 4.52]
in third trimester, unadjusted	2.00 [0.80 - 4.96]
in third trimester, adjusted	**2.70 [1.00 - 7.29]**

Odds ratios and 95% confidence interval for the association between prenatal GU infection, flu (overall and by trimester), and fever (overall, by trimester, and intrapartum only) exposure for unadjusted and fully adjusted models (adjusted for child sex, maternal age, birth year, maternal smoking, education, marital status, race, gestational age). All exposure information was obtained by maternal self report 24-72 hours after delivery; with the exception of intrapartum fever > 38C, which was extracted from labor and delivery records.

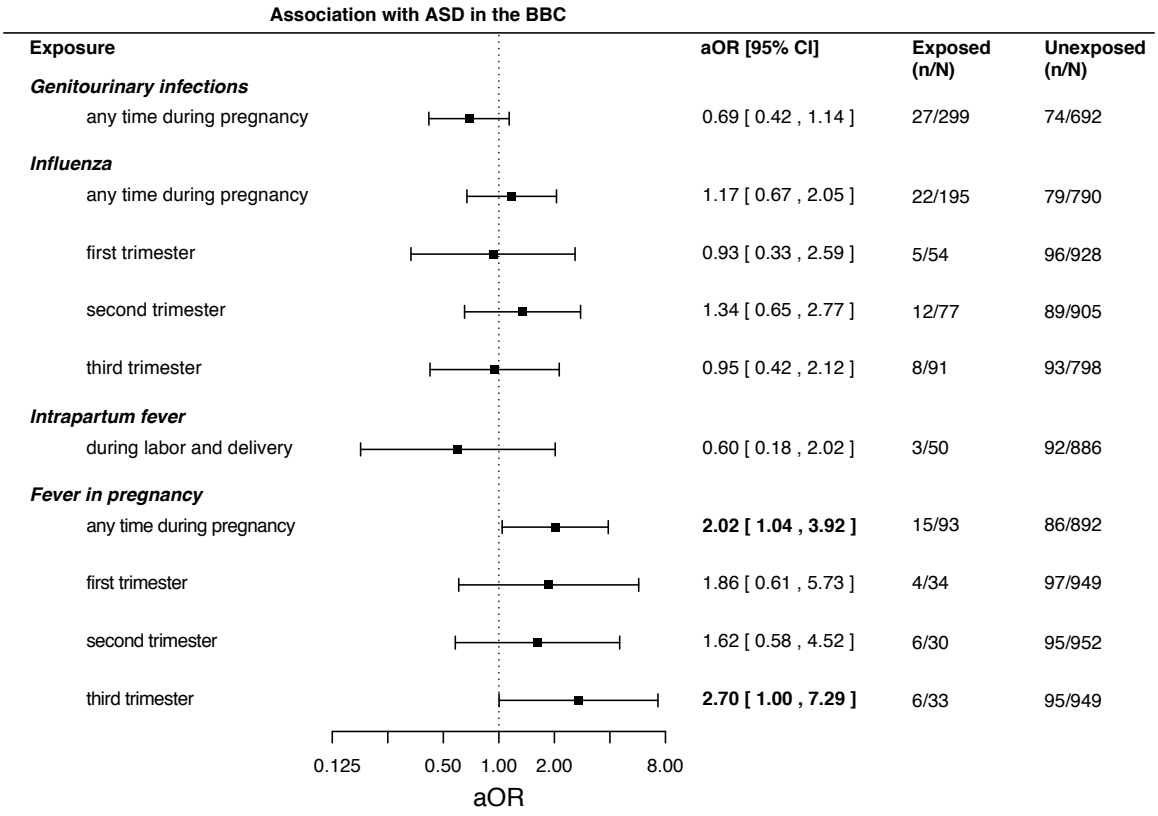


Figure 2.1: Forest plot showing adjusted odds ratio (OR) and 95% confidence intervals for the association between prenatal GU infection, flu (overall and trimester-specific), and fever (overall, trimester-specific, and intrapartum) and autism in the Boston Birth Cohort (BBC). Sample sizes are shown for those who were exposed or unexposed for each variable, with the number with the ASD outcome (n) over the total who were in that category, including both ASD cases and neurotypical controls (N).

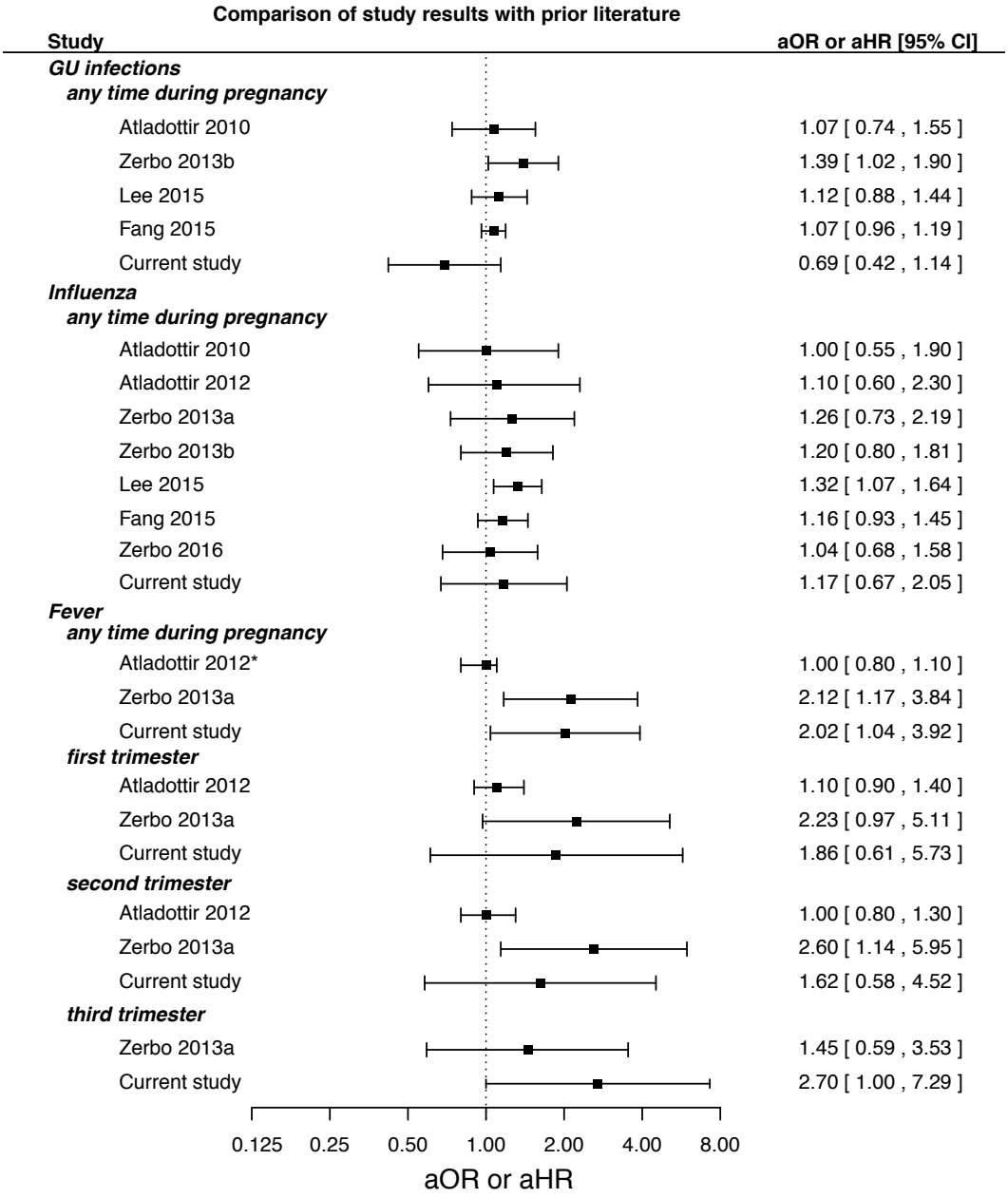


Figure 2.2: Forest plot comparing the BBC results to previously reported results. The plot shows effect estimates and 95% confidence intervals for the association between infection or fever (at any point during pregnancy and by trimester) and autism.

2.4.3 Prenatal Exposure to Influenza and ASD Risk

We assessed risk for ASD among children prenatally exposed to influenza at any point during gestation or during a specific trimester, including trimesters 1, 2, and 3. We observed no association between prenatal influenza exposure at any point during gestation and risk of ASD in either unadjusted (OR of 1.14 [95% confidence interval 0.69 – 1.89]) or adjusted analyses (aOR 1.17 [0.67 – 2.05]; Figure 2.1 and Table 2.7). Furthermore, neither adjusted nor unadjusted analyses showed an association between ASD risk and prenatal influenza exposure specific to any trimester (Figure 2.1 and Table 2.7). This is consistent with prior studies that reported no association between influenza infection, at any time during pregnancy, and risk of the offspring developing ASD (Figure 2.2).

2.4.4 Fever during pregnancy is associated with increased ASD risk in offspring

No association between exposure to fever at any time during pregnancy and risk of ASD was found in unadjusted analyses (OR 1.80 [0.99 - 3.27]; Table 2.7). However, a significant association between child ASD diagnosis and maternal fever at any time during her pregnancy was found after adjustment for child sex, maternal age, child birth year, maternal smoking status, maternal education, marital status, maternal race, and gestational age (aOR 2.02 [1.04 – 3.92]; see Figure 2.1 and Table 2.7). We also observed a significant association between maternal fever during the third trimester of pregnancy and child ASD diagnosis (aOR 2.70 [1.00 – 7.29]; Figure 2.1 and Table 2.7). No association between child ASD status and exposure to maternal fever during the first

(aOR 1.86 [0.61 – 5.73]) or second trimesters (aOR 1.62 [0.58 – 4.52]) was found (Figure 2.1 and Table 2.7).

Despite different covariate choices, the association between prenatal fever exposure and ASD was consistent in magnitude and direction across all regression models tested (Table 2.8). However, because we observed differences in the strength of association among the adjusted models we also estimated the fever association using propensity score matching. Eighty-nine individuals exposed to fever were matched to 540 unexposed individuals; good balance on all covariates was achieved after propensity score matching (see Table 2.9 for characteristics of mother-child pairs by prenatal fever exposure status and Figure 2.3). Consistent with our regression models, matching by propensity score showed a similar association between gestational exposure to maternal fever and ASD risk (Table 2.8).

Because fever and influenza may be correlated, as fever is a potential symptom of influenza, we sought to further tease apart the relationship between prenatal exposure to fever/flu and ASD. We generated a categorical variable with four levels: neither flu nor fever (reference category; n = 753); fever only (n = 36); flu only (n=137); or both flu and fever (n = 56). We performed a logistic regression with ASD as the outcome and this derived categorical variable as the exposure, with adjustment for additional covariates as in our other models. No exposure category was significantly associated with ASD risk when compared to the reference category: fever (aOR 2.11 [0.75 – 5.94], p = 0.16), flu only (aOR 0.94 [0.47 – 1.90], p = 0.87), or fever and flu (aOR 2.05 [0.90 – 4.65], p = 0.09).

Table 2.8: Comparison of effect size and significance for association between ASD and prenatal fever exposure using different analytic models

Analytic model	Covariates									OR	*P*
	Maternal					Child					
	educa-tion	marital status	race	smoking	age	GA	sex	birth year	LBW[a]		
1										1.80 [0.99 - 3.27]	0.053
2										1.72 [0.93 - 3.19]	0.085
3										1.80 [0.96 - 3.38]	0.068
4										1.86 [0.97 - 3.58]	0.062
5*										**2.02 [1.04 - 3.92]**	**0.037**
6										**1.98 [1.02 - 3.86]**	**0.045**
7										**1.95 [1.01 - 3.77]**	**0.046**
8										1.89 [0.995 - 3.59]	0.051
9	propensity score matching[b]									1.54 [0.80 - 2.95]	0.195

GA, gestational age; LBW, low birth weight
Bold values represent statistically significant ORs

[a] low birth weight is defined as weighing less than 2500 grams at birth
[b] conditional probability of exposure to fever modeled as a function of child sex, maternal age, birth year, maternal smoking, education, marital status, race, gestational age
* All main text references to "fully adjusted model" refer to this model

Table 2.9: Characteristics of mother-child pairs in the Boston Birth Cohort (BBC), by status of prenatal expose to fever (any time during pregnancy)

	Unexposed (n=892)	Exposed (n=93)	*p*
ASD diagnosis, *n* (%)			0.075
no	806 (90.4)	78 (83.9)	
yes	86 (9.6)	15 (16.1)	
Maternal age[a], M (SD)	28.55 (6.53)	27.63 (6.59)	0.196
Child year of birth, M (SD)	2006.22 (3.53)	2005.83 (3.40)	0.312
Education, *n* (%)			0.055
Elementary school	34 (3.8)	6 (6.5)	
Secondary school	221 (24.8)	19 (20.4)	
High school/GED	316 (35.4)	23 (24.7)	
Some college	186 (20.9)	29 (31.2)	
College degree and above	135 (15.1)	16 (17.2)	
Marital status, *n* (%)			0.483
Married	316 (35.4)	29 (31.2)	
Not Married	576 (64.6)	64 (68.8)	
Race or Ethnicity[b], *n* (%)			<0.001*
Black	568 (63.7)	37 (39.8)	
White	75 (8.4)	9 (9.7)	
Hispanic	181 (20.3)	38 (40.9)	
Asian	16 (1.8)	5 (5.4)	
Other	52 (5.8)	4 (4.3)	
Maternal smoking[c], *n* (%)			0.336
Never	765 (84.8)	77 (82.8)	
Some	48 (5.4)	3 (3.2)	

Continuous	88 (9.9)	13 (14.0)	
Child sex, *n* (%)			0.609
Female	500 (56.1)	49 (52.7)	
Male	392 (43.9)	44 (47.3)	
Gestational age, mean (SD)[d]	38.21 (3.09)	38.28 (2.49)	0.831
Birth weight, *n* (%)			0.635
>2500 grams	695 (77.9)	75 (80.6)	
<2500 grams	197 (22.1)	18 (19.4)	

ASD, autism spectrum disorder; BBC, Boston Birth Cohort; M, mean; SD, standard deviation

[a] Maternal age at time of delivery

[b] Black includes self reported Black, African American, Haitian, Cape Verdean, and Caribbean race and ethnicities. Asian includes Asian and Pacific Islander races. The Other category includes individuals with a mixed or other racial background.

[c] Never smokers were defined as mothers with no history of smoking 6 months prior to conception or during pregnancy; some smoking includes mothers that smoked at some point in the window of 6 months prior to conception and delivery but did not smoke throughout that window; continuous is defined as mothers that smoked starting 6 months prior to and throughout pregnancy.

[d] Defined by sonogram

* Denotes statistically significant

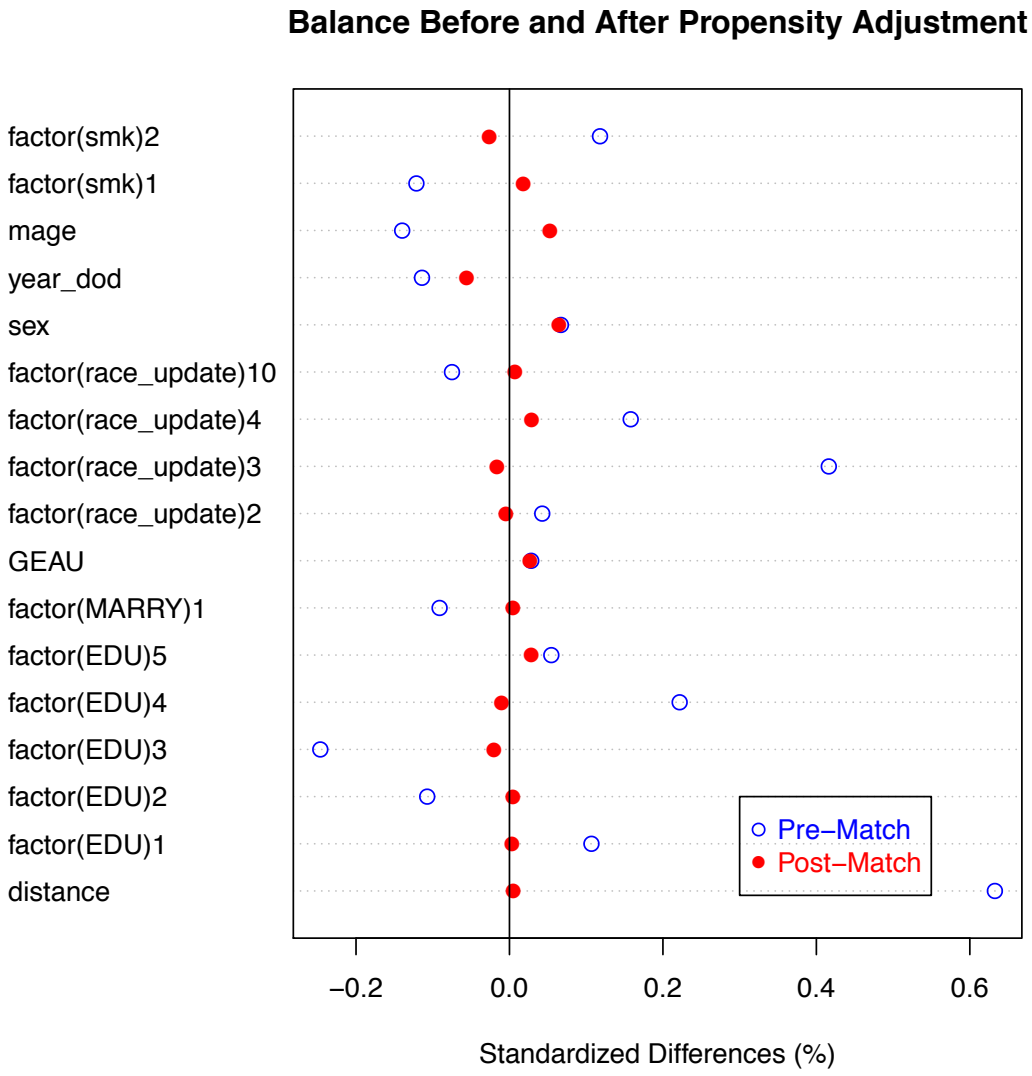


Figure 2.3: Love plot demonstrating covariate balance before and after propensity score matching.

To further test the association between exposure to fever during gestation and risk of ASD, we also performed a sensitivity analysis for potential outcome misclassification. We used a more stringent ASD case definition of 2 or more 299 ICD-9-CM diagnoses. When the logistic regression models were repeated for ASD cases with at least 2 diagnosis codes (n=82) and neurotypical controls (n=884), the unadjusted OR for ASD risk after fever exposure at any time during pregnancy was 1.95 [1.03 – 3.68] and the adjusted OR was 2.11 [1.03 – 4.32]. Using a more stringent ASD case definition did not change the estimated strength of the association between prenatal fever exposure and ASD risk or its significance.

Finally, we compared our findings to two prior studies that examined the association between prenatal exposure to maternal fever and ASD risk. Zerbo et al. saw an association between fever and ASD risk at any time during pregnancy, consistent with the current study (Figure 2.2) (Zerbo 2013a). For exposure to maternal fever during pregnancy (prior to 32 weeks) or specifically during trimester 1 or 2, Aladottir et al. did not find an association with ASD risk (Aladottir 2012).

2.4.5 Intrapartum fever is not associated with ASD risk

Because of the association we observed between maternal fever during the third trimester and the risk of ASD in her child, we wanted to examine intrapartum fever exposure. This variable represents fever in the peripartum period (labor and delivery) only, as opposed to the third trimester. No association was seen between intrapartum fever and child ASD status in unadjusted (OR 0.52 [0.16 - 1.69], Table 2.7) or adjusted

analyses (aOR 0.60 [0.18 - 2.02]; Figure 2.1 and Table 2.7). Because there were only four individuals with both intrapartum fever and fever at any prior point during pregnancy (as measured by a structured interview based on the standardized postpartum questionnaire), we were not able to assess the joint effect.

2.4.6 Effect of exposure misclassification

Because our study relies on self-reported exposure data collected shortly after delivery, we were concerned about the possibility of exposure misclassification biasing our results. We used two strategies for quantitative bias estimation with our three main exposure categories (infection, flu any time during pregnancy, fever any time during pregnancy): a probabilistic or Monte Carlo sensitivity analysis (MCSA) and bootstrapping, as implemented in the R package *episensr* (Lash et al. 2009). Briefly, these methods use estimates of the sensitivity and specificity of the exposure test (in this case, a structured interview based on a questionnaire) to reassign false negatives and false positives in a 2x2 table. The corrected 2x2 table is then used to calculate a crude odds ratio; the distribution of this OR over many replicates, either with randomly drawn sensitivity and specificity and the actual study sample per the MCSA approach or on a particular bootstrapped sample, is then used to generate an empirical corrected OR estimate and 95% confidence interval.

For the MCSA, we chose to model the prior probability of the sensitivity for both cases and controls as a uniform distribution between 0.5 and 1, while the specificity was a uniform distribution between 0.912 and 1. We note here that because the correction

method for exposure misclassification involves subtracting false positive subjects, we are unable to assess for dramatic overreporting of exposure (very low specificity) because it would result in negative cells in the corrected 2x2 table. Because the exposure data is collected prior to the ASD outcome, we are less concerned about differential exposure misclassification by ASD status. We thus set the sensitivity and specificity draws for the cases and controls to be tightly correlated (r = 0.9) and ran 10,000 replicates. For GU infection, the OR corrected for systematic exposure misclassification and random error was 0.78 [95% CI 0.43 - 1.38]. For flu, the OR corrected for systematic exposure misclassification and random error was 1.21 [0.67 - 2.17]. For fever, the OR corrected for systematic exposure misclassification and random error was 2.68 [1.17 - 9.59].

In our second approach, we first performed a bootstrap resampling of the dataset (10,000 bootstraps) and then conducted a simple sensitivity analysis for misclassification on each bootstrap. This allows us to model different ways that bias could affect the sensitivity and specificity of the test for exposure (Table 2.10). We believe there is low risk of differential exposure misclassification because of the study's prospective data collection, and so the sensitivity and specificity for ASD cases and controls are the same.

Table 2.10: Results of simple sensitivity analysis for misclassification with 10,000 bootstraps for each exposure-bias combination

Exposure	Sensitivity	Specificity	Adjusted OR [95% CI]
GU infection			
Under-reporting	0.5	0.99	0.73 [0.33 - 1.57]
	0.6	0.99	0.77 [0.40 - 1.49]
	0.7	0.99	0.79 [0.45 - 1.44]
	0.8	0.99	0.81 [0.48 - 1.40]
Good test performance	0.95	0.95	0.80 [0.46 - 1.44]
Over-reporting	0.95	0.90	0.77 [0.39 - 1.65]
Under- and over-reporting	0.6	0.90	0.72 [0.31 - 1.76]
Flu			
Under-reporting	0.5	0.99	1.21 [0.58 - 2.54]
	0.6	0.99	1.18 [0.62 - 2.29]
	0.7	0.99	1.17 [0.65 - 2.17]
	0.8	0.99	1.16 [0.67 - 2.09]
Good test performance	0.95	0.95	1.19 [0.63 - 2.35]
Over-reporting	0.95	0.90	1.27 [0.52 - 3.47]
Under- and over-reporting	0.6	0.90	1.30 [0.48 - 4.00]
Fever			
Under-reporting	0.5	0.99	2.07 [0.96 - 4.67]
	0.6	0.99	2.01 [0.98 - 4.33]
	0.7	0.99	1.96 [0.98 - 4.09]
	0.8	0.99	1.93 [1.00 - 3.96]
Good test performance	0.95	0.95	**2.77 [1.02 - 8.14]**
Over-reporting	0.95	0.92	1.80 [0.99 - 3.27]
Under- and over-reporting	0.6	0.92	**9.43 [1.47 - 86.28]**

2.5 Discussion

We examined the prospective relationship between prenatal maternal immune activation and Autism Spectrum Disorder (ASD) risk in an understudied, predominantly urban minority population. Our results support previous suggestions of maternal immune

activation as a risk factor for ASD, and specifically implicate fever. Our results do not provide evidence for an association between exposure to genitourinary infections or influenza during gestation and later diagnosis of ASD. However, we did observe a significant association between prenatal exposure to fever at any time during pregnancy as well as during the third trimester, and ASD.

This result is consistent with most prior studies showing ASD risk is not related to prenatal exposure to genitourinary or influenza infection (Figure 2.2). Only two prior studies have specifically examined prenatal exposure to fever, at any point in pregnancy or during specific trimesters, and ASD risk, with conflicting results. Our result showing increased ASD risk associated with fever (without regards to length of the febrile episode) is consistent with Zerbo et al. (Figure 2.2) (Zerbo et al. 2013) but not Atladottir et al. (Atladottir et al. 2012). The prenatal exposure definition used by Atladottir et al. included only data through week 32; however, our BBC study and the Zerbo et al. report used fever exposure for the entire pregnancy period (week 1 to birth). Thus, the Atladottir et al. definition did not include a large portion of the third trimester, shown to be significantly associated with ASD in our BBC data (Figures 2.1 and 2.2). Although the overall findings between Zerbo et al. and our study were consistent, their study identified a significant association with fever in the second trimester while we observed a significant association with fever in the third trimester. It is possible that this difference could be due to recall bias and/or exposure misclassification due to the difference in study design.

Our study sample was derived from an enriched risk cohort, where children were initially recruited with oversampling for preterm birth, a known ASD risk factor. BBC exposure data was collected at birth, prior to ASD diagnoses, allowing for a relatively short recall time and prospective analysis for ASD. In contrast, Zerbo et al. (Zerbo et al. 2013) used a retrospective case-control design in which the exposure data was ascertained up to 60 months after birth (longer recall) and after ASD diagnosis. Additionally, our study population has a different racial and economic make-up (Table 2.1) than the population in Zerbo et al., which was approximately 50% white (Zerbo et al. 2013).

Our findings were specific to fever, which could indicate that fever itself contributes to ASD risk. There is evidence that exposure to fever during pregnancy can lead to several different suboptimal developmental outcomes including oral clefts, neural tube defects, and congenital heart defects (Dreier et al. 2014). In support of this, Zerbo et al. (Zerbo et al. 2013) found an attenuation of the fever-ASD development association with the use of antipyretics such as acetaminophen. However, rather than fever itself being in the causal ASD pathway, it is also possible that fever is merely acting as a marker of a specific infection that is associated with increased ASD risk but that is not captured by our questionnaire data. The BBC prenatal infection exposure data is collected soon after birth with a structured questionnaire designed to ascertain exposures in a reliable manner. Nonetheless, the questionnaire is retrospective with respect to the exposure and it is possible that mothers remember fever better than infections generally. Finally, it is possible that the pathology of ASD begins during gestational development

and leads to increased maternal infections and fever through immunocompromise, although there is little evidence to support this.

We observed relatively large confidence intervals in our study compared to prior work (Figure 2.2). This is likely due to the relatively small number of individuals in our study and thus imprecision in these estimates. Because our study relied on the extraction of ASD outcomes from electronic medical records, specifically ICD-9-CM codes, there may be some outcome misclassification. However, we would not expect any outcome misclassification to be differential by exposure status, and a sensitivity analysis using a more stringent ASD case definition continued to show a significant association between fever exposure at any time during gestation and ASD. Additionally, there is significant co-morbidity among the ASD cases in our sample for developmental delay; of our 116 ASD cases, 111 also have a diagnosis for a developmental delay (ICD-9-CM codes between 315.0 and 315.9). Prior research in a similar EMR data set has found this to be true among validated ASD cases (Dodds et al. 2009), and so we believe there is limited outcome misclassification.

We did not perform an adjustment for multiple comparisons. We examined ten total exposures (GU infections; flu, overall and by trimester; fever, overall and by trimester; and intrapartum fever), and the significant association we found between prenatal exposure to fever and ASD risk would not survive a conservative Bonferroni correction. Additional research with larger samples is necessary to clarify the role of prenatal exposure to fever in the development of ASD.

Future studies in urban, low-income, minority populations are needed to replicate our findings. In addition, future work to evaluate potential clinical interventions is needed. One previous study from the Danish National Birth Cohort did assess the association between anti-pyretic medications and ASD (Liew et al. 2016). Unexpectedly, the results showed an increased risk for ASD with hyperkinetic symptoms and maternal use of acetaminophen during pregnancy (Liew et al. 2016), which contradicts the results from Zerbo et al. To help resolve these differences in the literature, future studies should collect and evaluate information on prenatal exposure to fever, including (1) timing during pregnancy, (2) severity, or degree of temperature elevation, (3) duration of fever, (4) the probable etiology, including specific infections, and (5) use of anti-pyretics, including timing and dose.

There has been limited prior research on the relationship between prenatal fever exposure and ASD risk, particularly among underrepresented and minority populations. Our findings expand upon past work, and provide evidence supporting prenatal exposure to fever as a risk factor for ASD in an urban, low income, minority population, adding to our knowledge of a highly prevalent modifiable risk factor that may inform public health strategies for primary and secondary prevention.

Chapter 3: Developing methods utilizing Machine Learning and Latent Class Analysis to identify children with ASD in administrative health data

This chapter describes work currently in preparation for submission for peer review, with contributions from co-authors Rashelle Musci, M. Dani Fallin, Xiaobin Wang, Elizabeth A. Stuart, and Christine Ladd-Acosta.

3.1 Introduction

3.1.1 Electronic medical record-based research

The increase in the availability of EHR data has accelerated since the Health Information Technology for Economic and Clinic Health (HITECH) Act of 2009, which mandated the adoption and meaningful use of EHR across the US (Roden and Denny 2016). Until September 30, 2015, medical billing data in the United States used the International Classification of Diseases (ICD) code, ninth edition, clinical modification (ICD-9-CM) (O'Malley, Cook et al. 2005). The ICD classification system is frequently used for a diverse range of purposes including insurance reimbursement, health administration, and epidemiological research. It is necessary to find the best way to utilize these existing medical billing records. Electronic medical records are a potentially rich source of information for research on medical diagnoses, treatment, and disease course, but there are several challenges that need to addressed. In particular, a need for

improved sensitivity and specificity of case identification from electronic medical records has been identified by researchers working in a variety of fields.

The electronic MEdical Records & Genomics (eMERGE) network was created in 2007 by the National Human Genome Research Institute to establish the utility of electronic medical records in genetic research; this has led to the development of phenome-wide association studies (PheWAS) to explore multiple EMR phenotype-to-genetic variant associations (Denny, Bastarache et al. 2013, Crawford, Crosslin et al. 2014, Verma, Basile et al. 2016). eMERGE and PheWAS "phecodes" aggregate individual ICD-9-CM diagnoses into code groups and major disease concept paths, but largely rely on a simple algorithm based on the presence or absence of individual codes (Namjou, Marsolo et al. 2014) (https://phewascatalog.org). There may be additional information within the medical record that can be used to inform phenotype construction and improve it beyond current strategies.

We are interested in the ways that EMR, and specifically billing or administrative health data rather than free-text clinical notes, can be optimized for the use of autism researchers. Research on ASD risk factors increasingly utilizes large datasets derived from electronic medical records or medical claims data; ensuring the validity of ASD phenotypes derived from these records is essential (Dodds, Spencer et al. 2009). Because ASD is still a relatively rare disease, despite its increased prevalence in the United States, EMR may be particularly useful for autism research. A traditional prospective birth cohort of a general population sample would require massive recruitment to generate reasonable sample sizes with an expected ASD prevalence of approximately 1.5% among

offspring (Christensen, Baio et al. 2016). With EMR, we can generate case-control samples from hospital-based data and dramatically improve sample size and power for examining ASD risk factors.

It is often standard to use the simple presence or absence of an ICD code to identify cases and controls in administrative health databases. While the US currently uses ICD-10-CM, and ICD-10 has been used in most other countries since the mid 1990s (Bowman, Cleland et al. 2015), most US-based research studies use historical data and will continue to have a need to identify ASD cases from ICD-9-CM codes. ASD diagnosis codes in ICD-9-CM are based on the DSM-IV definitions of disease: 299.0 (autistic disorder), 288.8 (Asperger's syndrome), and 299.9 (pervasive developmental disorder not otherwise specified, PDD-NOS). Some studies may require that a patient be assigned an ASD code twice or by a pediatric neurodevelopmental specialist, while others allow for any provider to assign the code even once (Dodds, Spencer et al. 2009, Wang and Leslie 2010, Peacock, Amendah et al. 2012, Zerbo, Qian et al. 2013, Connolly, Anixt et al. 2016; also see above, Chapter 2, section 2.3.2 Analytic Sample & Outcome Classification).

Requiring the presence of two ASD-category ICD-9-CM codes, rather than just a single ASD code, may optimize the sensitivity and specificity (Coleman, Lutsky et al. 2015). However, there are social factors, including hospital resources, that may influence the likelihood of a child with ASD receiving a diagnosis (Mazumdar, Winter et al. 2013). We also know that electronic medical records are subject to both random and systematic error: this may be due to the quality of information available to the provider,

communication errors between patient and provider or between providers, clinician experience and skill with the diagnosis, coder training and experience, facility practices, or strategies that may manipulate what is billable for a particular visit, including misspecification, unbundling, and upcoding (O'Malley, Cook et al. 2005). Error in EMR, driven both by coding practices and social factors related to receiving the appropriate diagnosis, complicates attempts to conduct epidemiologic research on ASD risk factors.

Straightforward use of these ICD-9-CM codes—even when requiring two diagnoses of ASD at separate visits—may not be the best way to derive valid phenotypes for epidemiologic research. Prior work showed a sensitivity of only 69.3% compared to a gold standard clinical diagnosis when the presence of one ASD code was used to define case status; sensitivity dropped significantly, to 36.9%, when two ASD codes were required to be present in the medical record to define a case (Dodds, Spencer et al. 2009). Either strategy for identifying children with ASD in an administrative health database would result in a number of false negative cases, which could bias the results of an association analysis.

One particular way to approach this problem would be to utilize the full array of development, behavioral, and language-related diagnoses that a child may receive from a medical provider. We know that false negative ASD cases are likely to carry diagnoses of related behavioral and developmental disorders, as well as certain psychiatric and medical conditions that are known to co-occur with ASD (Dodds, Spencer et al. 2009, Levy, Giarelli et al. 2010, Doshi-Velez, Ge et al. 2014). It may be possible to use these other diagnoses to improve the sensitivity (identify false negative ASD cases) and

specificity (identify false positives, or children incorrectly coded as having ASD) of case identification in electronic medical records.

3.1.2 Co-occurring conditions in Autism Spectrum Disorder

There is an extensive body of literature examining the prevalence and patterns of co-occurring conditions in individuals with Autism Spectrum Disorder, over both secular time and developmental age, by sex, and by intellectual disability status (Close, Lee et al. 2012, Kohane, McMurry et al. 2012, Stacy, Zablotsky et al. 2014). The majority of children with ASD have at least one non-ASD developmental diagnosis (83%), and a sizable fraction have at least one other psychiatric diagnosis (10%) or neurologic diagnosis (16%) (Levy, Giarelli et al. 2010). The non-ASD developmental diagnoses in this study included ADHD, language disorder, learning disability, intellectual disability, nonverbal learning disabilities, and sensory integration disorder; the clustering of these diagnoses along with ASD is so consistent that it is recommended children with any one of them be comprehensively evaluated (Gillberg 2010). Additionally, children and young adults with ASD are more likely than hospital-based controls to have epilepsy, schizophrenia, inflammatory bowel disease, other bowel disorders, CNS/cranial anomalies, diabetes mellitus type I, muscular dystrophy, and sleep disorders (Kohane, McMurry et al. 2012). The frequency of co-occurring conditions may actually be increased in children who meet ASD criteria but have yet to receive the appropriate diagnosis, indicating that co-occurring conditions complicate the diagnostic odyssey for

ASD and may be particularly useful in identifying false negative ASD cases in EMR (Levy, Giarelli et al. 2010).

It is possible that reproducible clusters of specific co-occurring conditions exist in children with autism; these clusters may reflect disease etiology or subtype. For example, a child with ASD and gastrointestinal disease is likely to also have sleep problems; and the severity of behavioral problems can be predicted by the number of co-occurring medical conditions a child with ASD has (Aldinger, Lane et al. 2015). The observation of clustering of particular symptoms and co-occurring conditions in individuals with ASD has begun to be formalized with statistical methods that can perform unsupervised clustering on the basis of comorbidity patterns. One study using hierarchical clustering of subgroups of children with ASD found a group characterized by seizures, a group with multisystem disorders including GI disorders and infections, and a group with psychiatric disorders (Doshi-Velez, Ge et al. 2014). There are other potential approaches that have yet to be applied to ASD.

3.1.3 Random Forests

Data mining techniques may prove useful in understanding patterns in child development as captured by administrative health data. One particular method is Random Forests, an extension of Classification and Regression Trees or CART, in which observations or "features" are partitioned to predict a particular outcome (Breiman 1984). Random Forests (RF) is an ensemble method, where multiple trees are constructed and then allowed to vote on the optimal partitioning of the data for outcome prediction

(Breiman 2002). RF can be performed with classification trees, where the predicted outcome is class membership; or with regression trees, in which the predicted outcome is some value of a continuous variable. One advantage of RF is that information about the predictive utility or importance of individual features can be obtained from the model; this may assist in developing an algorithm that "boosts" case identification on the basis of co-morbid conditions.

RF has previously been shown to work well in disease prediction from electronic medical records (Khalilia, Chakraborty et al. 2011). It has been successfully used to predict chronic diseases such as diabetes (Zheng, Xie et al. 2017), hypertension (Khalilia, Chakraborty et al. 2011), and fibromyalgia (Emir, Masters et al. 2015), though its use is not yet widespread and translation into clinical management has been limited (Clifton, Niehaus et al. 2015). To our knowledge, RF has yet to be applied to medical claims data for the purpose of distinguishing normal and abnormal child neurodevelopment, though two other machine learning techniques, Support Vector Machine and Generalized Low Rank Modeling, have previously been applied to ASD cohorts to distinguish phenotypic subtypes (Lingren, Chen et al. 2016, Schuler, Liu et al. 2016), and RF has been applied to text phrases from developmental evaluations (Maenner, Yeargin-Allsopp et al. 2016).

Here we use the results of ASD screening performed in the Boston Birth Cohort's Children's Health Study (CHS) to construct a regression tree with random forests, as well as the presence of a 299 group code to construct a classification tree. We determine the variable importance of specific ICD-9-CM codes present in the medical record for predicting performance on an ASD screening questionnaire or the presence of a 299

diagnosis. We then use the most important ICD-9-CM codes in a latent class analysis of the CHS cohort.

3.1.4 Latent Class Analysis

Latent class analysis (LCA) is a probabilistic method for clustering individuals based on their observed characteristics or "indicators;" individuals are grouped based on an unknown (latent) variable—class membership—that explains the correlation between their observed characteristics (Collins and Lanza 2010). This technique has been applied to administrative health datasets to identify patients with asthma (Prosser, Carleton et al. 2008), systemic autoimmune diseases (Bernatsky, Lix et al. 2011), and sepsis (Shahraz, Lagu et al. 2014). Here we use LCA to identify clusters of children with ASD and distinguish disease severity and co-morbidity; the observed characteristics are the ICD-9-CM diagnoses recorded in the child's medical record.

3.1.5 Study Hypothesis

We hypothesized that LCA could be used in the BBC to identify children with probable ASD on the basis of related behavioral and developmental diagnoses, as well as frequent medical co-morbidities such as epilepsy. Latent class analysis requires choosing the relevant observations or indicators; in this case, ICD-9-CM diagnosis or diagnosis categories served as indicators. We used machine learning, specifically Random Forests, to best classify the CHS population and determine which ICD-9-CM diagnosis codes best identify the children with abnormal ASD screening results. These predictive ICD-9-CM

diagnoses were then used as indicators in the generation of latent classes. Our methods may reveal additional cases who do not yet have an ASD diagnosis in their medical record, or identify the children who are most severely affected.

3.2 Methods

3.2.1 Boston Birth Cohort (BBC)

The Boston Birth Cohort is a prospective birth cohort, established as the Molecular Epidemiology of Preterm Delivery in 1998 with the recruitment of women delivering at the Boston Medical Center (BMC), with oversampling for preterm birth (Wang, Zuckerman et al. 2002). Women with a singleton live birth at the BMC are eligible for recruitment, with exclusions for IVF, multiple gestations, chromosomal abnormalities, major birth defects, and preterm deliveries due to maternal trauma. They are contacted by the research team 24-72 hours after birth. Written informed consent is obtained and mothers are interviewed using a standardized questionnaire (Maternal at Postpartum Questionnaire) to gather information about her pregnancy. Maternal and infant medical records are reviewed and data on ultrasound findings, placental pathology, laboratory reports, pregnancy complications, labor and delivery course, and birth outcomes are extracted. Biospecimens from mother and child are obtained and banked for future research. The Children's Health Study (CHS) was established in 2002 to enable follow-up of child developmental outcomes. Research staff approach mothers at

scheduled pediatric clinic visits and obtain informed consent for the CHS follow-up study. There are no exclusions for participation in CHS if the mother consents. Medical records are then obtained, including type and date of visits, clinical diagnosis, medications, and growth and development parameters.

The BBC serves as an enriched-risk cohort for Autism Spectrum Disorders. At its establishment, mother-child pairs were oversampled for preterm births, a known risk factor for ASD. Consequently there is an estimated ASD prevalence of 4.0% in the entire CHS cohort, compared to the general population prevalence of 1.1-1.5% (Christensen, Baio et al. 2016). The BBC thus serves as an excellent resource for the prospective study of the causes, risk factors, and disease course of ASD, which is otherwise still relatively rare in the general population. Information about child development is primarily available through the Boston Medical Center's administrative health data. These records include information on visit date and location, insurance, and the patient's medical diagnoses. However, a growing subset of children have been given an ASD screening questionnaire, the Social Communication Questionnaire (SCQ) (Rutter 2003, Chandler, Charman et al. 2007); and a quantitative measure of autistic traits, the Social Responsiveness Scale (SRS) (Constantino, Davis et al. 2003).

We used R package *tableone* (https://CRAN.R-project.org/package=tableone) for the construction of tables demonstrating population characteristics in the BBC. Descriptive statistics for categorical variables were obtained with the function chisq.test() with continuity correction, and the function oneway.test() for continuous variables with

an assumption of equal variance. Annotation of ICD-9-CM codes was performed with the R package *icd* (https://CRAN.R-project.org/package=icd).

3.2.2 Standard ASD identification algorithm

A simple ICD-9-CM code-based algorithm for identifying children with ASD (cases) and children with neurotypical development (controls) was carried out in the CHS (Table 3.1). As is standard in the literature, ASD cases were identified as any one occurrence of ICD-9-CM codes 299.0 (autism), 298.8 (Asperger's syndrome), and 299.9 (pervasive developmental disorder not otherwise specified), including both active (.00) and residual (.01) disease states. We classified individuals as neurotypical controls if they were never diagnosed with any of the following conditions: ASD, attention deficit hyperactivity disorder (ADHD), intellectual disability (ID), developmental delay (DD), oppositional defiant disorder (ODD) or other "emotional disorders of childhood," conduct disorder (CD), or congenital anomalies. We also determined the number of times a child was given a particular diagnosis and the clinic type (specialist or general pediatric) in which ASD was diagnosed.

Table 3.1: ICD-9-CM code based definitions for ASD cases and typically developing controls in the Boston Birth Cohort, 2003 – 2015

	ICD-9-CM codes	N
ASD case definition		120
Inclusion criteria (any one of these codes):	299.0, 299.01, 299.8, 299.81, 299.9, 299.91	
Neurotypical control definition		1033
Exclusion criteria (any one of these codes):		
ADHD	314.0 − 314.9	
Conduct Disorder	312.0 − 312.9	
Emotional disturbances of childhood or adolescence including Oppositional Defiant Disorder	313.0 − 313.9	
Developmental Delay	315.0 − 315.9	
Intellectual Disability	317 − 319	
Congenital Anomalies	740 − 759.9	

3.2.3 Random Forests

There is a widely used R package for implementing Random Forests called *randomForest* (Wiener 2002) (https://CRAN.R-project.org/package=randomForest). This package allows the construction of both classification and regression trees using the function randomForest(), and the ranking of input features based on the extent to which their removal from the model increases predictive error, using the function importance() and varImpPlot() for plotting. We used three sets of input features in the construction of regression trees for the prediction of child score on the SCQ: (1) all codes included, including 299 group codes; (2) all codes except 299 group codes, which are removed; and

(3) all codes except for 299, V, and E codes, which are removed. V and E codes are supplemental ICD-9 codes; V codes are used to classify non-illness related visits to a medical professional, such as a well-child visit, while E codes are used to classify injury or poisoning. We then used the second feature matrix (all codes except for 299 codes) to generate a classification tree for predicting the presence of 299.00 (autism, current state) based on the presence of other ICD-9-CM codes. Variable importance was extracted to determine which ICD-9-CM diagnoses provided the most information for predicting the score on the SCQ, or for predicting the presence of a 299.00 diagnosis. We used the R package *icd* (https://CRAN.R-project.org/package=icd) for the annotation of ICD-9-CM codes.

3.2.4 Latent Class Analysis

LCA can be implemented with a dedicated software program Mplus (https://www.statmodel.com), which is the standard in the field. Observed characteristics of the sample or "indicators" are used to identify latent subgroups within a particular population. One limitation of LCA is that you are restricted to a relatively small number of indicators in the construction of latent classes; while RF can handle feature matrixes with thousands of input features, LCA requires careful choice of not more than 10-15 indicators. Here we used as LCA indicators the ICD-9-CM codes identified by Random Forests as being most relevant to outcome prediction. We thus ran analyses in Mplus (with option Type=Mixture, and increasing random starts to ensure model convergence) of four different sets of indicators (3 from regression trees, and 1 from the classification

tree); within each indicator set, iterations with different class sizes were run and then model fit statistics (Bayesian Information Criteria (BIC), sample size adjusted BIC, Lo-Mendell-Rubin) compared to determine the class size of the best solution.

3.3 Results

3.3.1 Sample Description

CHS electronic medical records (covering the time span from 1 October 2003 to 30 September 2015, the last day in the US before the transition to ICD-10) include 118,939 pediatric inpatient, outpatient, and emergency room records from 2,992 children after removing records from siblings of index children and removing duplicate records. Each child contributes on average 39.8 visits to the dataset, with a range of 1 to 463 visits. Demographic information about children in the CHS was determined by linking pediatric EMR records to demographic information from postpartum maternal questionnaires, which could be done for 2932 children out of the 2992 children in the CHS cohort (Table 3.2).

Table 3.2: Characteristics of mother-child pairs in the Boston Birth Cohort (BBC)'s CHS, 2003 - 2015

		Overall (n= 2932)
Maternal age[a], M (SD)		28.5 (6.5)
Child year of birth, M (SD)		2006.4 (3.7)
Education, *n* (%)		
	Elementary school	128 (4.6)
	Secondary school	684 (24.5)
	High school/GED	1002 (35.8)
	Some college	611 (21.8)
	College degree and above	372 (13.3)
Marital status, *n* (%)		
	Married	928 (33.2)
	Not Married	1866 (66.8)
Race or Ethnicity[b], *n* (%)		
	Black	1932 (65.9)
	White	208 (7.1)
	Hispanic	616 (21.0)
	Asian	46 (1.6)
	Other	130 (4.4)
Maternal smoking[c], *n* (%)		
	Never	2276 (81.4)
	Some	210 (7.5)
	Continuous	310 (11.1)
Child sex, *n* (%)		
	Female	1447 (49.4)
	Male	1485 (50.6)
Gestational age, mean (SD)[d]		37.7 (3.5)

Birth weight, *n* (%)		
	>2500 grams	2063 (73.3)
	<2500 grams	751 (26.7)

M, mean; SD, standard deviation

[a] Maternal age at time of delivery

[b] Black includes self reported Black, African American, Haitian, Cape Verdean, and Caribbean race and ethnicities. Asian includes Asian and Pacific Islander races. The Other category includes individuals with a mixed or other racial background.

[c] Never smokers were defined as mothers with no history of smoking 6 months prior to conception or during pregnancy; some smoking includes mothers that smoked at some point in the window of 6 months prior to conception and delivery but did not smoke throughout that window; continuous is defined as mothers that smoked starting 6 months prior to and throughout pregnancy.

[d] Defined by sonogram

3.3.2 Standard ASD identification algorithm

Using simple ICD-9-CM based ASD case and neurotypical control definitions, we identified 120 ASD cases and 1,033 neutrotypical controls from the full CHS cohort (n=2992). We classified individuals as neurotypical controls if they were never diagnosed with any of the following conditions: ASD, attention deficit hyperactivity disorder (ADHD), intellectual disability (ID), developmental delay (DD), oppositional defiant disorder (ODD) or other "emotional disturbances of childhood," conduct disorder (CD), or congenital anomalies (Table 3.1). Likely because of the oversampling for preterm birth in the design of the BBC and CHS, there was a large proportion of children with competing developmental and behavioral diagnoses that disqualified them from serving as a typically developing control.

Twenty-six subjects (out of 120) identified as ASD cases with a relaxed definition (≥ 1 ASD diagnosis) only had a single 299 group diagnosis; though notably, all of these children were also diagnosed with other developmental diagnoses that are known to commonly co-occur with ASD, including developmental delay, delayed milestones, or abnormal physiological development, speech or language disorders, attention deficits with or without hyperactivity, and behavioral or conduct disorders. In eight of these subjects, the ASD diagnosis was made by a specialist, either a developmental behavioral pediatrician (n=5), pediatric neurologist (n=2), or child psychiatrist (n=1).

3.3.3 Social Communication Questionnaire

Once a week, trained research staff pull upcoming appointments at the Boston Medical Center for research participants. Study personnel will meet the family in the waiting room of their scheduled appointment, and ask them to complete the Social Communication Questionnaire and/or Social Responsiveness Scale. Through 9 October 2016, 888 total SCQs were completed. If a child was administered the SCQ more than once, we retained the first score and discarded replicates. After removing duplicate exams and linking scores to pediatric electronic medical records through a family-level study ID, 771 unique subjects had both EMR and SCQ data, while 745 have EMR, SCQ, and demographic information (Table 3.3). Compared to other children in the CHS who did not receive the SCQ, children with a recorded SCQ were less likely to be low birth weight ($p = 0.018$), more likely to identify as black and less likely to identify as white (p

= 0.002), younger (born on average in 2007, rather than 2005, $p < 0.001$), with mothers who were less likely to have smoked during their pregnancy ($p = 0.019$).

We used the raw score of the SCQ; in our dataset, this ranged from 0 to 24 (out of 40 yes/no questions), with a mean score of 6.67 and standard deviation of 4.26 (Figure 3.1). While some studies use an SCQ cutoff of 15 to indicate a positive screen for ASD, other studies have used a cutoff of 11 to improve sensitivity (Eaves, Wingert et al. 2006, Schendel, Diguiseppi et al. 2012). Ninety-six children had an SCQ score greater than 11, while 32 had an SCQ greater than 15.

Table 3.3: Characteristics of mother-child pairs with child SCQ data in the CHS, 2003 - 2015

		Overall (n= 745)
Maternal age[a], M (SD)		28.9 (6.6)
Child year of birth, M (SD)		2007.9 (3.2)
Education, *n* (%)		
	Elementary school	32 (4.6)
	Secondary school	151 (21.8)
	High school/GED	248 (35.8)
	Some college	167 (24.1)
	College degree and above	95 (13.7)
Marital status, *n* (%)		
	Married	220 (31.8)
	Not Married	471 (68.2)
Race or Ethnicity[b], *n* (%)		
	Black	522 (70.1)
	White	35 (4.7)
	Hispanic	156 (20.9)
	Asian	6 (0.8)
	Other	26 (3.5)
Maternal smoking[c], *n* (%)		
	Never	573 (83.3)
	Some	58 (8.4)
	Continuous	57 (8.3)
Child sex, *n* (%)		
	Female	355 (47.7)
	Male	390 (52.3)
Gestational age, mean (SD)[d]		37.3 (3.8)

Birth weight, *n* (%)		
	>2500 grams	485 (69.8)
	<2500 grams	210 (30.2)

M, mean; SD, standard deviation

[a] Maternal age at time of delivery

[b] Black includes self reported Black, African American, Haitian, Cape Verdean, and Caribbean race and ethnicities. Asian includes Asian and Pacific Islander races. The Other category includes individuals with a mixed or other racial background.

[c] Never smokers were defined as mothers with no history of smoking 6 months prior to conception or during pregnancy; some smoking includes mothers that smoked at some point in the window of 6 months prior to conception and delivery but did not smoke throughout that window; continuous is defined as mothers that smoked starting 6 months prior to and throughout pregnancy.

[d] Defined by sonogram

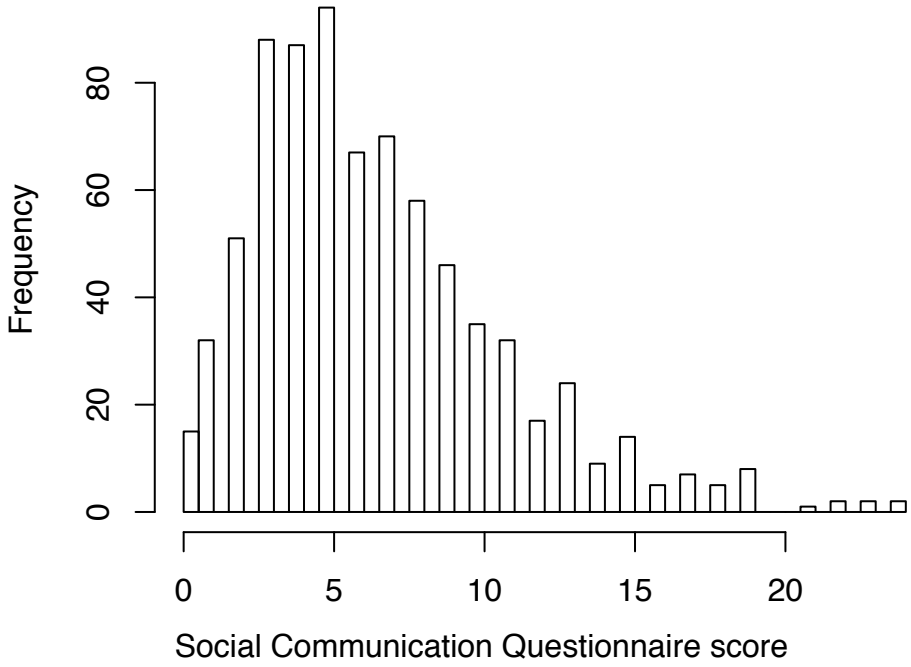


Figure 3.1: Distribution of SCQ scores in the Children's Health Study (n = 771).

3.3.3 Random forests

In the 771 patients with SCQ data, we extracted the unique ICD-9-CM codes that were associated with any of their medical visits in their electronic medical records and used the presence or absence of a single ICD-9-CM code to generate a dichotomous yes/ no variable for each child. A child was given a '1' if their record contained a particular code at any point in time, and a '0' if the code was never associated with their medical record. This generated 2355 unique features per child, and a feature matrix with 771 rows and 2355 columns. We then used Random Forests as implemented in the R package *randomForest* to identify the specific ICD-9-CM codes most predictive of the continuous SCQ score with regression trees. Parameters included ntree=500 and importance=TRUE so that we could later assess the variable importance of each ICD-9-CM code-derived feature. We used three different sets of ICD-9-CM codes for the regression trees: (1) all codes included, including 299 group codes (features = 2355); (2) all codes except 299 group codes (features = 2351) and (3) all codes except for 299, V, and E codes (features = 1957). Next, we used the same set of features from our second regression tree (all codes except 299 group codes, features = 2351) to predict the presence or absence of the ICD-9-CM code 299.00, which is "autism, current state." As ICD-9-CM is based on DSM-IV diagnostic criteria, this diagnosis reflects the most severe form of ASD. In the subset of children with SCQ scores, 50 of them had at least one diagnosis of 299.00 in their medical record.

The first set of features (all codes included, including 299 group codes) best explained the variance in the SCQ scores (Table 3.4). After each model, we used the

varImpPlot() and importance() functions to identify the ICD-9-CM codes most predictive of SCQ score (Figures 3.2 - 3.5; Tables 3.5 - 3.8). The variable importance plots show the model performance when each feature is removed from the model; while the model performance is complex and can rely on the network of predictors, the increase in mean squared error and impact on node purity when a feature is removed from the model provides some measure of its function in the model.

To determine if the predictive utility of these diagnoses was specific to ASD, we also built a classification tree to predict diagnosis with 466.19 ("Acute bronchiolitis due to other infectious organisms"). We found that the ICD-9-CM diagnoses most predictive of 466.19 do not overlap with those codes predictive of either SCQ score or the presence of a 299.00 diagnosis (Figure 3.6, Table 3.9).

Table 3.4: Random Forests model performance for regression and classification trees

	Mean of squared residuals	**% Variance explained**
Regression trees		
set 1: all codes	15.22	16.21
set 2: no 299 codes	16.13	11.2
set 3: no 299, V, or E codes	16.05	11.65
	Out Of Bag estimate of error rate	
Classification tree	6.49%	

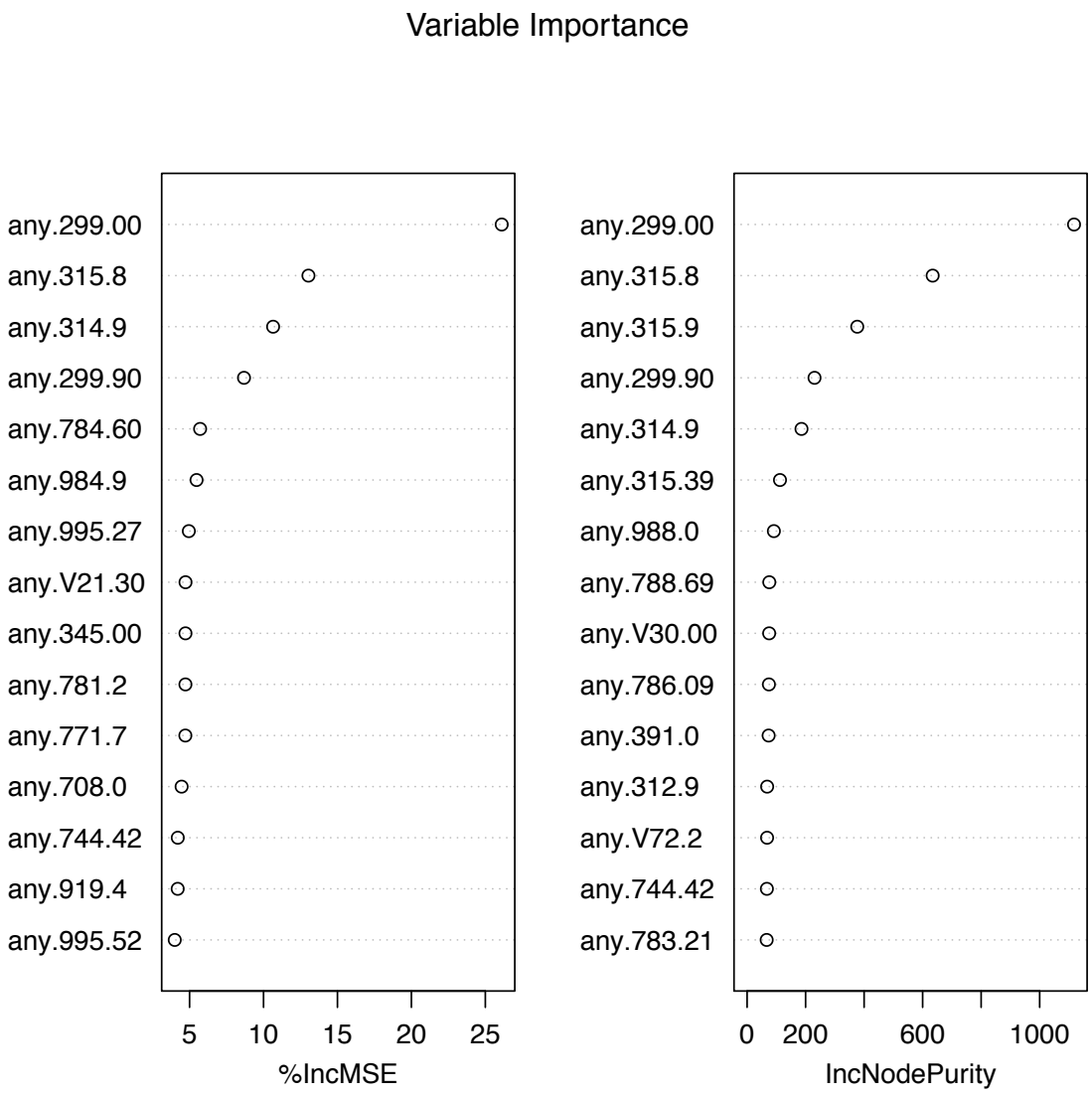


Figure 3.2: Variable importance plot; all codes included, including 299

Table 3.5: ICD-9-CM codes with highest variable importance, based on percent increase in mean squared error when each is removed from the feature matrix derived from all codes (including 299)

299.00	**Autistic disorder, current or active state**
315.8	**Other specified delays in development**
314.9	**Unspecified hyperkinetic syndrome**
299.90	**Unspecified pervasive developmental disorder, current or active state**
784.60	**Symbolic dysfunction, unspecified**
984.9	**Toxic effect of unspecified lead compound**
995.27	**Other drug allergy**
V21.30	**Low birth weight status, unspecified**
345.00	**Generalized nonconvulsive epilepsy, without mention of intractable epilepsy**
781.2	**Abnormality of gait**
771.7	**Neonatal Candida infection**
708.0	**Allergic urticaria**
744.42	**Branchial cleft cyst**
919.4	**Insect bite, nonvenomous, of other, multiple, and unspecified sites, without mention of infection**
995.52	Child neglect (nutritional)
327.23	Obstructive sleep apnea (adult)(pediatric)
V06.1	Need for prophylactic vaccination and inoculation against diphtheria-tetanus-pertussis, combined [DTP] [DTaP]
703.8	Other specified diseases of nail
9.1	Colitis, enteritis, and gastroenteritis of presumed infectious origin
V49.9	Unspecified problems with limbs and other problems

Top 14 predictive features in bold. These are used as indicators in downstream Latent Class Analysis.

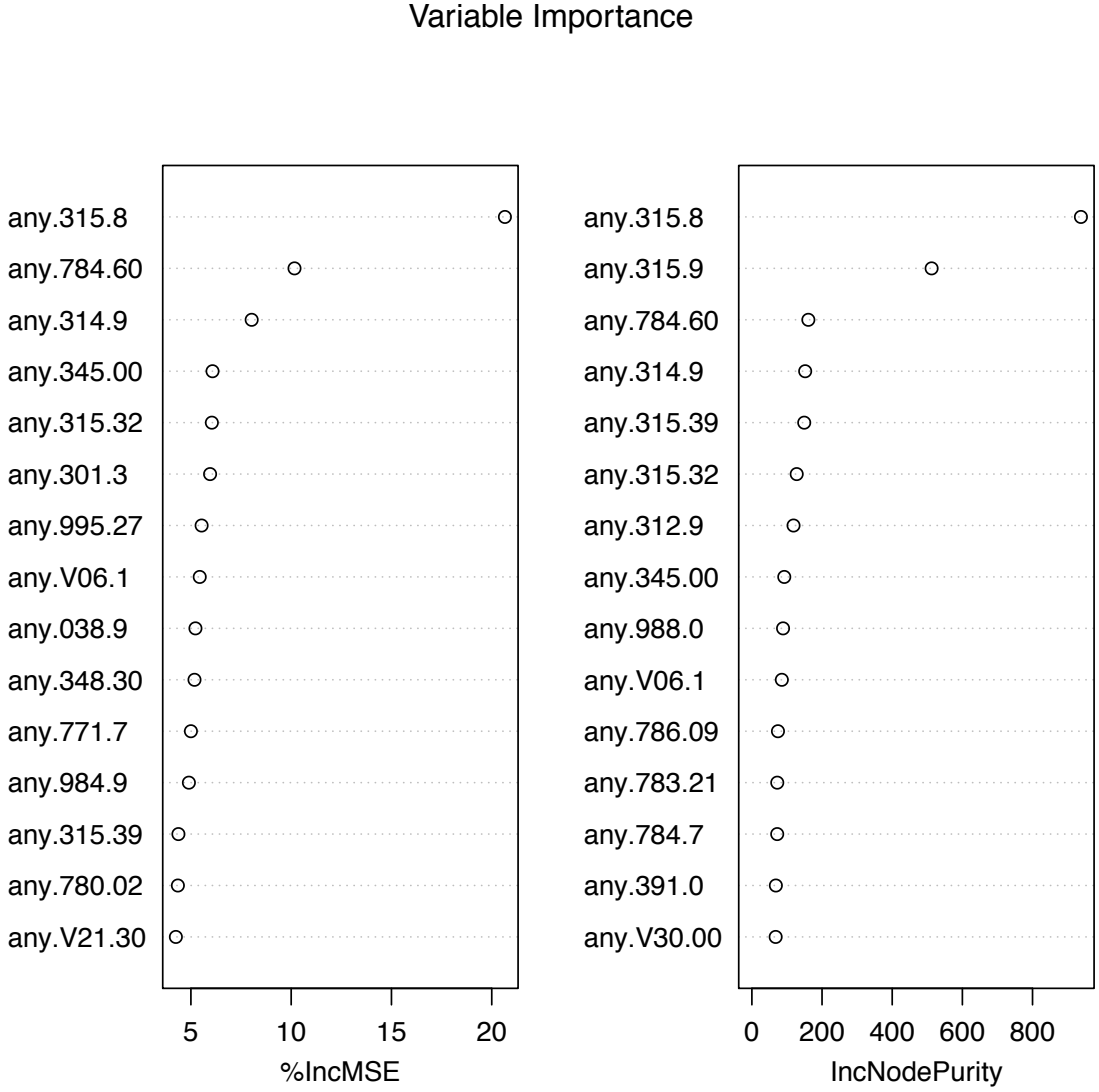


Figure 3.3: Variable importance plot; 299 codes removed from the predictors

Table 3.6: ICD-9-CM codes with highest variable importance, based on percent increase in mean squared error (feature matrix derived from codes excluding 299)

315.8	**Other specified delays in development**
784.60	**Symbolic dysfunction, unspecified**
314.9	**Unspecified hyperkinetic syndrome**
345.00	**Generalized nonconvulsive epilepsy, without mention of intractable epilepsy**
315.32	**Mixed receptive-expressive language disorder**
301.3	**Explosive personality disorder**
995.27	**Other drug allergy**
V06.1	**Need for prophylactic vaccination and inoculation against diphtheria-tetanus-pertussis, combined [DTP] [DTaP]**
38.9	**Unspecified septicemia**
348.30	**Encephalopathy, unspecified**
771.7	**Neonatal Candida infection**
984.9	**Toxic effect of unspecified lead compound**
315.39	**Other developmental speech or language disorder**
780.02	**Transient alteration of awareness**
V21.30	Low birth weight status, unspecified
781.2	Abnormality of gait
780.52	Insomnia, unspecified
744.42	Branchial cleft cyst
41.49	Other and unspecified Escherichia coli [E. coli]
779.84	Meconium staining

Top 14 predictive features in bold. These are used as indicators in downstream Latent Class Analysis.

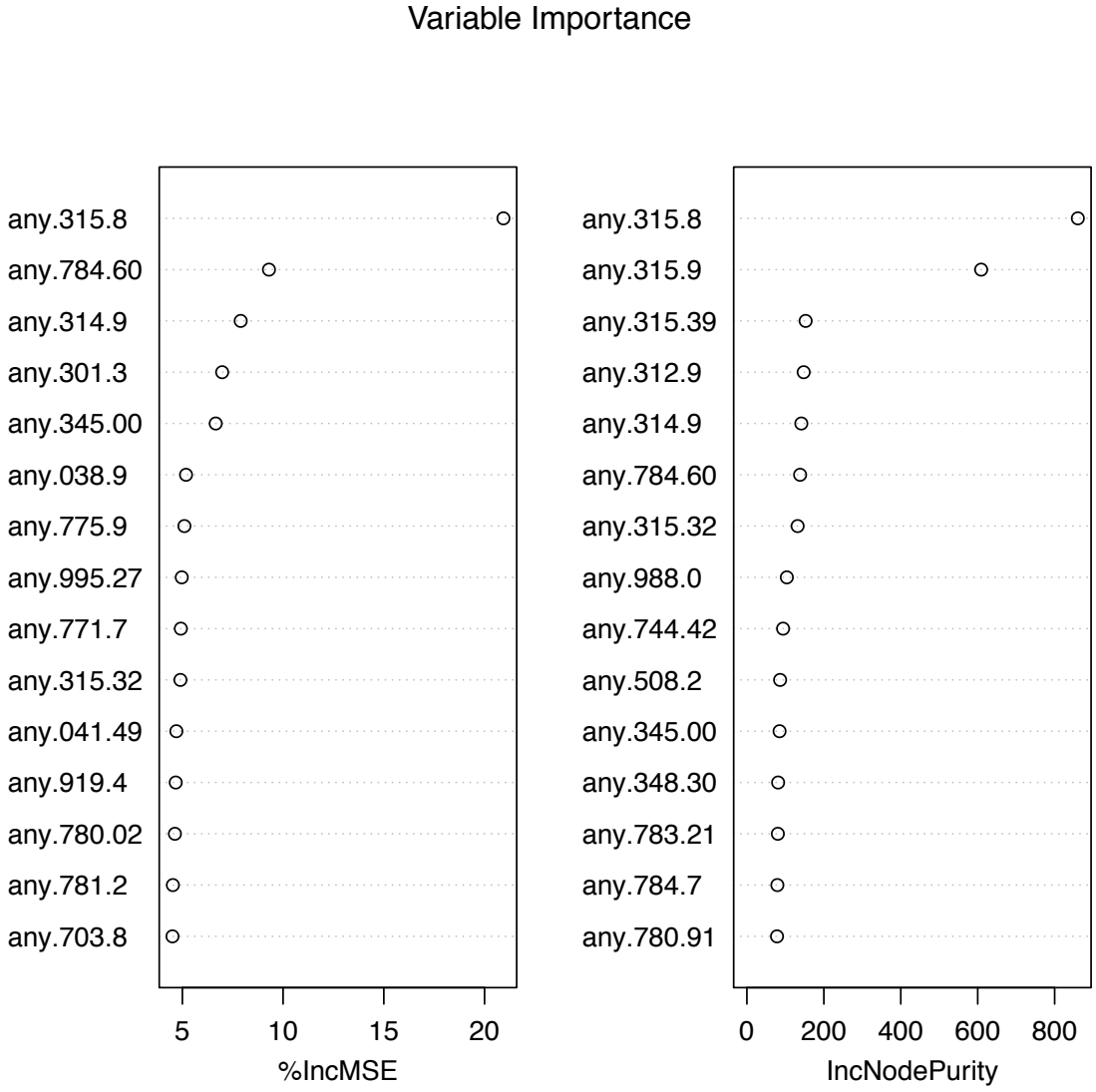


Figure 3.4: Variable importance plot; 299, V, and E codes removed

Table 3.7: ICD-9-CM codes with highest variable importance, based on percent increase in mean squared error (feature matrix derived from codes excluding 299, V, and E)

315.8	**Other specified delays in development**
784.60	**Symbolic dysfunction, unspecified**
314.9	**Unspecified hyperkinetic syndrome**
301.3	**Explosive personality disorder**
345.00	**Generalized nonconvulsive epilepsy, without mention of intractable epilepsy**
38.9	**Unspecified septicemia**
775.9	**Unspecified endocrine and metabolic disturbances specific to the fetus and newborn**
995.27	**Other drug allergy**
771.7	**Neonatal Candida infection**
315.32	**Mixed receptive-expressive language disorder**
41.49	**Other and unspecified Escherichia coli [E. coli]**
919.4	**Insect bite, nonvenomous, of other, multiple, and unspecified sites, without mention of infection**
780.02	**Transient alteration of awareness**
781.2	**Abnormality of gait**
703.8	Other specified diseases of nail
216.9	Benign neoplasm of skin, site unspecified
348.30	Encephalopathy, unspecified
779.84	Meconium staining
259.1	Precocious sexual development and puberty, not elsewhere classified
331.4	Obstructive hydrocephalus

Top 14 predictive features in bold. These are used as indicators in downstream Latent Class Analysis.

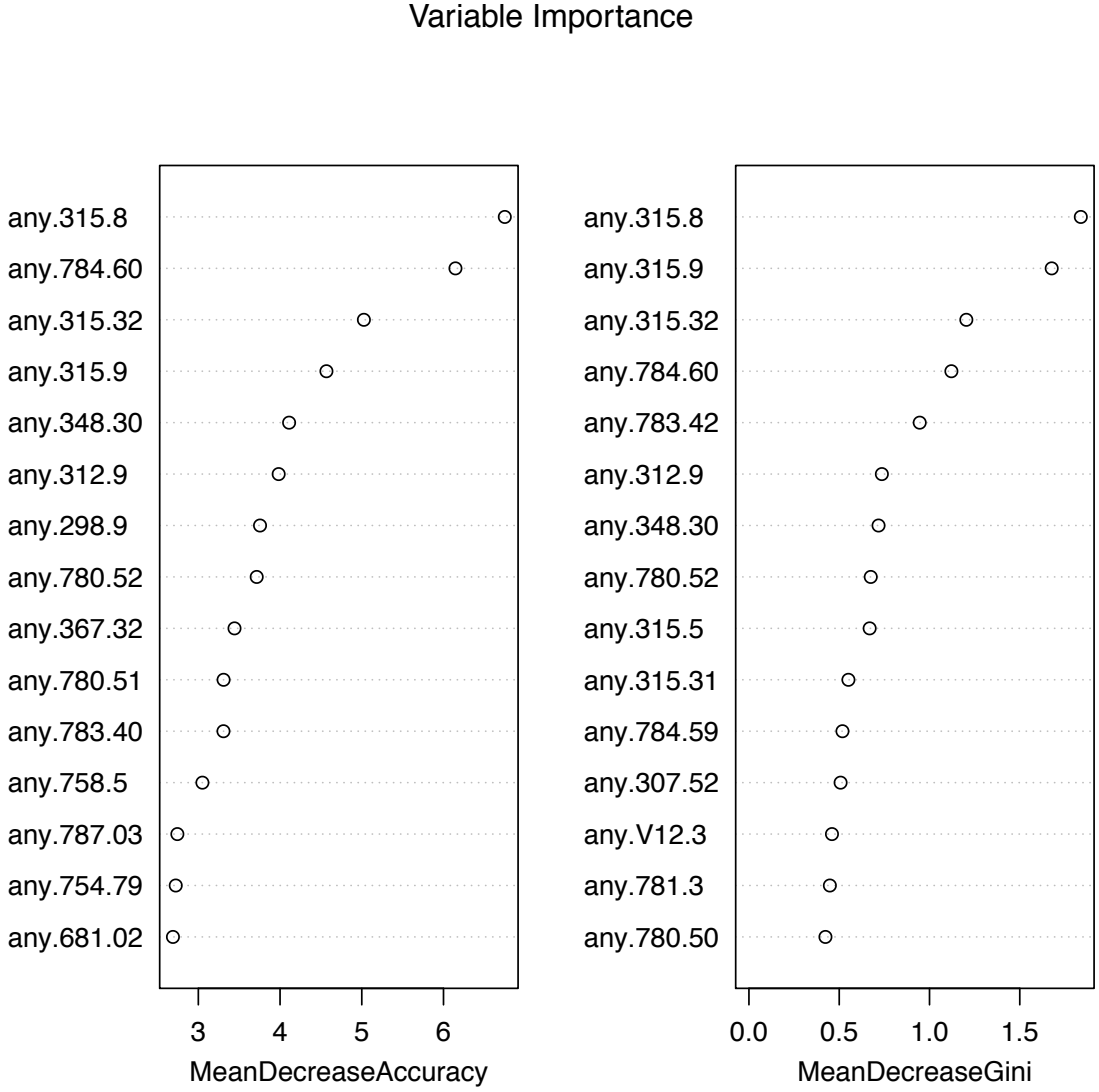


Figure 3.5: Variable importance plot of the classification tree: predict presence of 299.00 code based on the presence of other ICD-9-CM codes (V and E codes retained)

Table 3.8: ICD-9-CM codes with highest variable importance, based on percent increase in mean squared error (from the classification tree predicting 299.00 as outcome)

315.8	**Other specified delays in development**
784.6	**Symbolic dysfunction, unspecified**
315.32	**Mixed receptive-expressive language disorder**
315.9	**Unspecified delay in development**
348.3	**Encephalopathy, unspecified**
312.9	**Unspecified disturbance of conduct**
298.9	**Unspecified psychosis**
780.52	**Insomnia, unspecified**
367.32	**Aniseikonia**
780.51	**Insomnia with sleep apnea, unspecified**
783.4	**Lack of normal physiological development, unspecified**
758.5	**Other conditions due to autosomal anomalies**
787.03	**Vomiting alone**
754.79	**Other deformities of feet**
681.02	Onychia and paronychia of finger
314.9	Unspecified hyperkinetic syndrome
780.97	Altered mental status
752.65	Hidden penis
784.69	Other symbolic dysfunction
282.7	Other hemoglobinopathies

Top 14 predictive features in bold. These are used as indicators in downstream Latent Class Analysis.

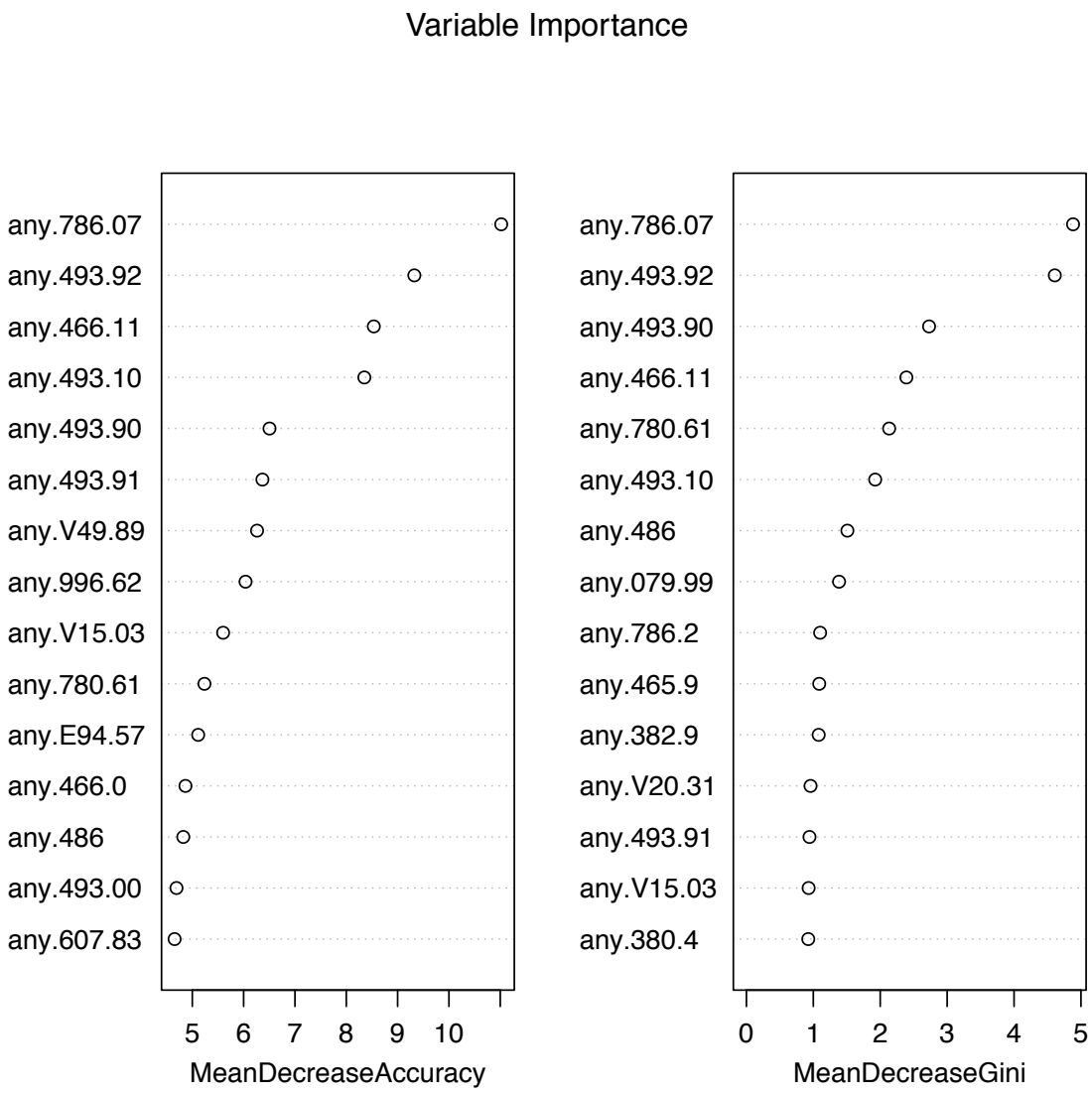


Figure 3.6: Variable importance plot for the prediction of a diagnosis with 466.19 (Acute bronchiolitis due to other infectious organisms). This serves as a negative control.

Table 3.9: ICD-9-CM codes with highest variable importance for the prediction of a diagnosis with 466.19 (Acute bronchiolitis due to other infectious organisms)

786.07	**Wheezing**
493.92	**Asthma, unspecified type, with (acute) exacerbation**
466.11	**Acute bronchiolitis due to respiratory syncytial virus (RSV)**
493.1	**Intrinsic asthma, unspecified**
493.9	**Asthma, unspecified type, unspecified**
493.91	**Asthma, unspecified type, with status asthmaticus**
V49.89	**Other specified conditions influencing health status**
996.62	**Infection and inflammatory reaction due to other vascular device,**
V15.03	**Allergy to eggs**
780.61	**Fever presenting with conditions classified elsewhere**
E94.57	**Antiasthmatics causing adverse effects in therapeutic use**
466	**Acute bronchitis**
486	**Pneumonia, organism unspecified**
493	**Extrinsic asthma, unspecified**
607.83	Edema of penis
480.1	Pneumonia due to respiratory syncytial virus
832.2	Nursemaid's elbow
E92.08	Accidents caused by other specified cutting and piercing instruments
V18.0	Family history of diabetes mellitus
701.4	Keloid scar

Top 14 predictive features for the presence of a 466.19 diagnosis in bold. These do not overlap with features predictive of SCQ score or the presence of a 299.00 diagnosis.

3.3.4 Latent class analysis

The codes with the highest importance criteria in the RF models were used as indicators in four separate latent class analyses (LCA) of the full dataset (n=2992). In the first analysis, we used 14 indicators from the full set of ICD-9-CM codes: dichotomous variables representing the presence or absence of any one diagnosis of 299.00 (Autistic disorder, current or active state), 315.8 (Other specified delays in development), 314.9 (Unspecified hyperkinetic syndrome), 299.90 (Unspecified pervasive developmental disorder, current or active state), 784.60 (Symbolic dysfunction, unspecified), 984.9 (Toxic effect of unspecified lead compound), 995.27 (Other drug allergy), V21.30 (Low birth weight status, unspecified), 345.00 (Generalized nonconvulsive epilepsy, without mention of intractable epilepsy), 781.2 (Abnormality of gait), 771.7 (Neonatal Candida infection), 708.0 (Allergic urticaria), 744.42 (Branchial cleft cyst), and 919.4 (Insect bite, nonvenomous, of other, multiple, and unspecified sites, without mention of infection).

LCA indicated a four class solution (Table 3.10), with a high entropy of 0.972, which demonstrates that assignment of the subjects to one of the four classes has little uncertainty. One class (93% of the sample) had a low probability of a 299 code or other developmental diagnoses (normal class), while one class (4% of the sample) had a high probability of carrying a 299 diagnosis (ASD-type class) (Figure 3.7). When compared to the ASD cases identified in the same dataset using standard definitions (n=120), the ASD-type class derived from LCA was smaller (n=113). There were 9 children who carried a 299 code but did not cluster with other children who did; and there were 2 children who did not carry a 299 code but were assigned to the ASD-type class on the

basis of other characteristics. While only a few of these children were tested with the SCQ, one of the children who clustered outside of the ASD class despite having a 299 diagnosis had an SCQ of 8, which is below even a relaxed cutoff of 11. The children in this group carried a number of diagnoses related to ADHD or hyperkinetic symptoms. Of the 113 children who clustered in the ASD class, 57 children received the SCQ, with a mean score of 12.4 (SD 5.4). The children who cluster outside of the ASD class may represent false positive ASD cases when using the standard ICD-9-CM-based identification, while children who cluster with the ASD class despite not having an ASD diagnosis of record may represent false negative ASD cases.

Table 3.10: Criteria evaluating the model fit for different class solutions, using the ICD-9-CM codes most predictive of SCQ score, including 299.00; this represents an effort to "boost" the utility of using a 299.00 code alone.

	LL	BIC	ssaBIC	# Free Par	LMR value	LMR p-value
Class 1	-3913.175	7938.402	7893.919	14		
Class 2	-3611.78	7455.668	7363.523	29	597.811	0
Class 3	-3586.73	7525.622	7385.817	44	49.687	0.0006
Class 4	-3572.007	7616.231	7428.765	59	29.203	0.0341
Class 5					13.803	0.3106

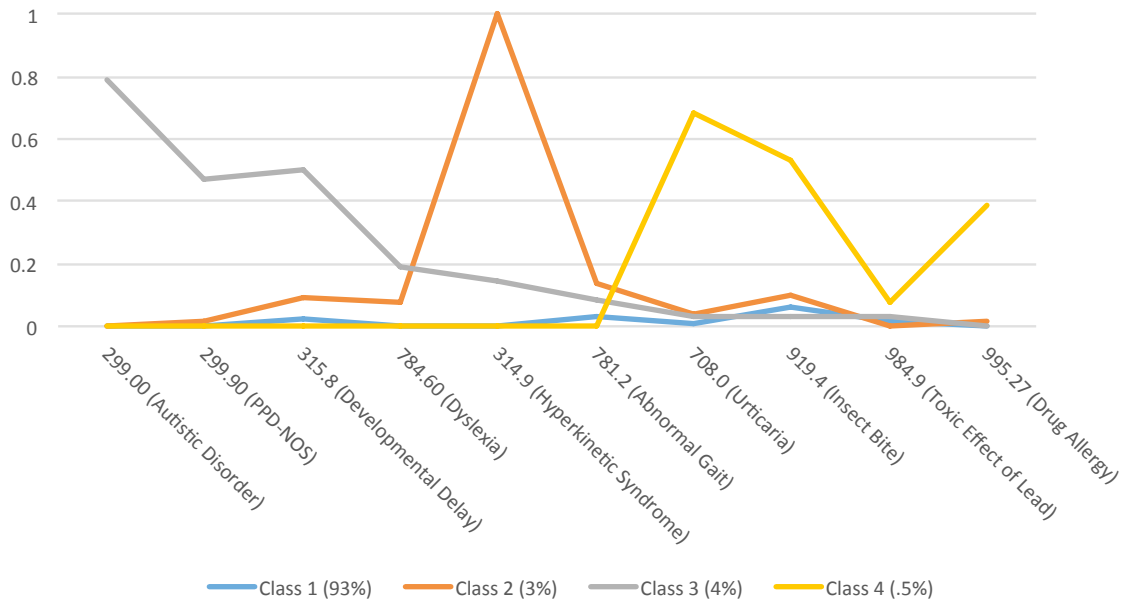


Figure 3.7: Illustrating the four class solution among the full set of 2992 children with electronic medical records from the BBC. ICD-9-CM diagnostic indicators were chosen based on their ability to predict SCQ score and have been pruned to those that distinguish the four classes.

When using the other two sets of indicators derived from the regression trees (299 codes removed from the predictors; and 299, V, and E codes removed), LCA indicated a 2 class solution. These were a normal class and smaller, atypical class with a higher prevalence of developmental diagnoses. However, the set of indicators derived from the classification tree predicting the presence of the most severe ASD diagnosis (299.00) indicated a four class solution (Table 3.11) with an entropy of 0.827. These indicators included 315.8 (Other specified delays in development), 784.6 (Symbolic dysfunction,

unspecified), 315.32 (Mixed receptive-expressive language disorder), 315.9 (Unspecified delay in development), 348.3 (Encephalopathy, unspecified), 312.9 (Unspecified disturbance of conduct), 298.9 (Unspecified psychosis), 780.52 (Insomnia, unspecified), 367.32 (Aniseikonia), 780.51 (Insomnia with sleep apnea, unspecified), 783.4 (Lack of normal physiological development, unspecified), 758.5 (Other conditions due to autosomal anomalies), 787.03 (Vomiting alone), and 754.79 (Other deformities of feet). Only 315.8 and 784.6 overlapped with the list of indicators used in the first LCA analysis. The 4 class solution indicated one typically developing class (75% of the population), and then three smaller classes with different types of developmental or medical diagnoses (Figure 3.8). Class 1 (15%) was distinguished by abnormal physiological development, Class 2 (3%) by encephalopathy and unspecified developmental delay in addition to abnormal physiological development, and Class 3 (7%) by psychosis and unspecified developmental delay.

Table 3.11: Criteria evaluating the model fit for different class solutions, using the ICD-9-CM codes most predictive of a 299.00 diagnosis

	LL	BIC	ssaBIC	# Free Par	LMR value	LMR p-value
Class 1	-7765.969	15643.99	15599.507	14		
Class 2	-7077.807	14387.72	14295.576	29	1364.956	0
Class 3	-7020.154	14392.471	14252.666	44	114.353	0
Class 4	-6988.127	14448.473	14261.007	59	63.524	0
Class 5*	-6973.036	14538.346	14303.219	74	29.415	0.2375

* model did not converge

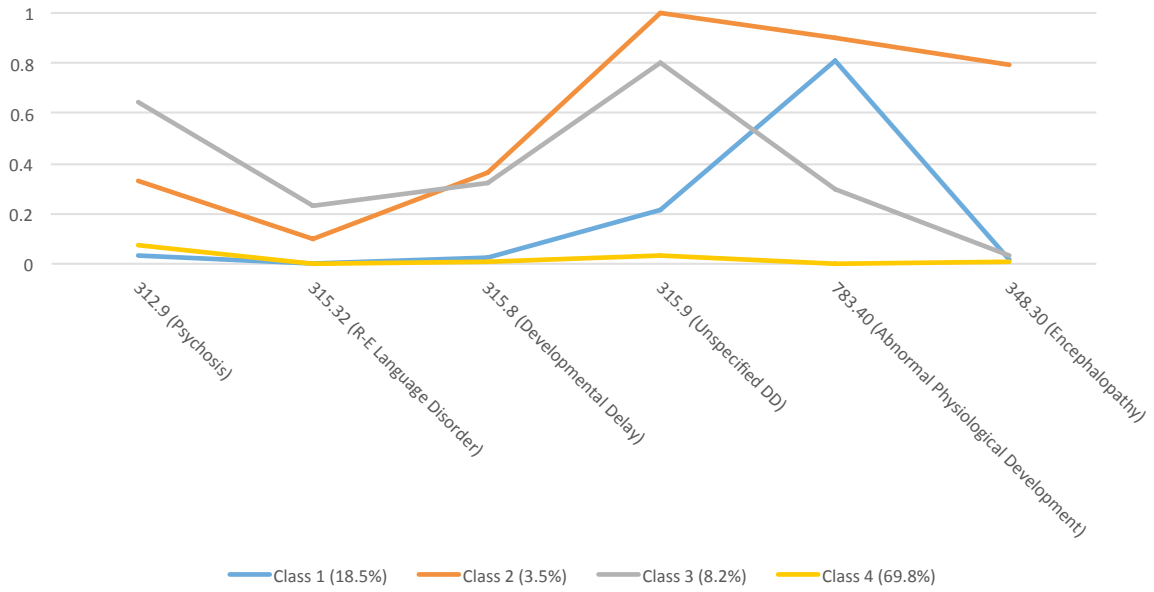


Figure 3.8: Illustrating the four class solution among the full set of 2992 children with electronic medical records from the BBC. ICD-9-CM diagnostic indicators were chosen based on their ability to predict a 299.00 diagnosis and have been pruned to those that distinguish the four classes.

3.4 Discussion

Here we were interested in ways to identify probable but undiagnosed ASD cases with latent class analysis in an administrative health dataset to assist epidemiologic research on ASD risk factors. We were also interested in the possibility of identifying homogenous, etiologically relevant clusters of children with abnormal neurodevelopment or neurodevelopmental disabilities. When we attempted to predict performance on an

ASD screening test, the Social Communication Questionnaire, with any ICD-9-CM code, including 299 and related codes, we found that in fact ASD diagnosis was the strongest predictor of SCQ performance. This is promising for two reasons: it suggests that the machine learning algorithm Random Forests (RF) is a valid method for understanding the patterns in EMR diagnoses that might predict a child's ASD status; and it provides a group of related ICD-9-CM diagnoses that may be used to "boost" or reinforce an ASD diagnosis, entirely based on information readily available in claims datasets. While RF in developmental or education records that have been parsed with natural language processing has promising utility for identifying children with ASD for surveillance (Maenner, Yeargin-Allsopp et al. 2016), this kind of data is generally less available, and with smaller sample sizes, than claims data.

We believe machine learning and Latent Class Analysis show promise in identifying children with typical or atypical neurodevelopment, leveraging the large amount of information available in the electronic medical records about a child's medical diagnoses over time. Because of the high degree of co-morbidity between ASD and other physical and psychiatric diagnoses, such as epilepsy or conduct disorder, we potentially can identify children who may have ASD but not yet carry a diagnosis in their EMR, or distinguish children on the basis of their disease severity.

This work was conducted in a single clinical cohort, using EMR from one medical institution. Further validation in a separate claims data set from another hospital or city will be necessary to understand the generalizability of the current findings. While our dataset covers twelve years of secular time, not all children contributed equal person-time

to the dataset; we have not yet taken into account child age or developmental stage into our models. While the CHS represents an opportunity to compare extensive EMR data with research-generated outcomes, we are also constrained by the smaller sample size available through the CHS. We hope to next extend these methods to an independent, larger administrative health dataset.

Future directions also include using random forests to predict scores in the Social Responsiveness Scale (SRS); SRS score itself can represent a broad, quantitative autistic phenotype that is distributed normally in the population (Constantino 2011). Currently, over 400 SRS questionnaires have been given to children in the BBC; as this sample size grows, we will begin to use the SRS data in addition to the SCQ.

3.5 Conclusion

Machine learning and latent class analysis show promise in improving ICD-9-CM-based ASD case identification. They may also help in identifying subgroups of children with ASD based on their clinical heterogeneity.

This work demonstrates a method for leveraging co-morbidity patterns in ASD for case and developmental subtype identification. It also raises the possibility that diagnostic patterns present in the medical record may improve early detection in clinical settings. Future extension of this work to larger samples in a variety of clinical settings will clarify its utility for improving research phenotypes and serving as a clinical tool.

Chapter 4: An epigenome-wide association study to detect epigenetic alterations reflecting prior exposure to infections *in utero*, amongst 2-5 year old children in the Study to Explore Early Development

4.1 Introduction

4.1.1 What is epigenetics?

Epigenetics is the study of DNA regulation that is mitotically heritable, or passed from one cell to its daughter cell (Allis and Jenuwein 2016). These are aspects of genetic regulation that do not change the underlying DNA code, but allow for differential gene expression in different cell types or for dynamic response to the environment. The expression of DNA can be regulated through histone modifications; histones are the protein cores that genomic DNA is wrapped around, like beads connected by string. Histones have a modifiable tail, where particular amino acids can be enzymatically altered to increase or decrease transcription of the local DNA sequence. In addition to histone modification, there is another well-studied form of epigenetic regulation: DNA methylation.

4.1.2 DNA methylation

In DNA methylation, a methyl group (or, less commonly, a hydroxymethyl group) added to a specific DNA nucleotide can influence the expression of a gene either locally or at-a-distance. In humans, most DNA methylation changes involve the enzymatic addition or removal of a methyl group to the carbon in the fifth position of the cytosine ring (resulting in 5-methylcytosine) at a cytosine-guanine dinucleotide (CpG site). These methyl marks on DNA can differ over time, across cell types, and in response to the environment (Ladd-Acosta 2015). In humans, DNA methylation is important in regulating development (Iurlaro, von Meyenn et al. 2017), has been implicated in cancer as either a part of the pathogenic pathway or as a biomarker (Juodzbalys, Kasradze et al. 2016), and reflects prior environmental exposures to toxicants (Huen, Yousefi et al. 2014) or even psychosocial stressors (Mehta, Klengel et al. 2013, Cecil, Smith et al. 2016).

4.1.3 Techniques for measuring DNA methylation

There are numerous strategies for detecting and localizing differential DNA methylation (Callinan and Feinberg 2006). Many of the techniques for detecting DNA methylation rely on the chemistry of cytosine and 5-methylcytosine (Olkhov-Mitsel and Bapat 2012, Kurdyukov and Bullock 2016). In the presence of sodium bisulfite, unmethylated cytosines will deaminate to uracil, while methylated cytosines will remain unchanged; downstream sequencing or array-based hybridization tools can determine which sites were converted to uracil, and hence were unmethylated, by comparison of the degenerate sequence with either the sequenced untreated input DNA or with genomic

references. If a researcher is interested in the DNA methylation at specific regions, a low throughput technique such as bisulfite pyrosequencing could be appropriate (Bassil, Huang et al. 2013). Genome-wide technologies include whole genome bisulfite sequencing. However, these strategies generally require more input DNA and are more expensive than array-based technologies, which have become the standard for large epidemiologic studies of epigenetic changes in humans.

4.1.4 Array-based DNAm measurement

DNA methylation can be assayed across the genome with an affordable platform, the Illumina Infinium HumanMethylation450 BeadChip methylation array ("450k platform" or "450k array"), which queries the degree of methylation at 485,512 genetic loci chosen to cover areas of biological interest, including genes, CpG islands, CpG island shores, FANTOM 4 promoters, predicted enhancers, and MHC regions (Bibikova, Barnes et al. 2011). This BeadChip methylation array evolved from an initial 27k platform (Bibikova, Le et al. 2009); a new platform with over 850,000 loci called MethylationEPIC is now available as well (Moran, Arribas et al. 2016). In the present study, we used the 450k array; to use this technology, input genomic DNA is first treated with sodium bisulfite, which converts unmethylated cytosines to uracil, as described above. The DNA is then denatured, neutralized, amplified, fragmented, and prepared for hybridization with the array BeadChips. Twelve samples are applied in separate wells to a single BeadChip, and hybridize to the 50mers covalently linked to the beads.

The 450k array uses two different types of probes to assay the degree of methylation at specific CpG sites across the genome. Type I probes actually use two bead types for a specific CpG locus; one is specific to the sequence expected if the CpG site is methylated, while another is specific to the sequence expected if it is unmethylated. Methylation state is detected by a single-base extension after hybridization of the probe to the target sequence; depending on whether the CpG site is methylated or unmethylated, that beadtype will fluoresce. In contrast, Type II probes have a single beadtype that can differentiate methylated and unmethylated CpG sites; it will fluoresce in the red channel if the CpG was unmethylated, while it will fluoresce in the green channel if the CpG was methylated. The complex chemistry and design of the 450k array presents challenges for the normalization and interpretation of 450k array data, as addressed below.

4.1.5 450k array and an epigenome-wide association study

Using the Illumina 450k array allows us to test for association of methylation at each measured locus with a phenotype of interest. This type of study is called an "epigenome-wide association study" or EWAS, in analogy to a genome-wide association study (GWAS)—though the 450k array is really genome-"scale," and not genome "wide," since by design the probes are densely clustered in some regions and absent in others (Chadwick, Sawa et al. 2015). In our study, we are interested in the detection of methylation differences that mark a history of prenatal exposure to maternal immune activation.

Previous work using this technology has demonstrated that prenatal insults can epigenetically alter offspring. For example, prenatal exposure to tobacco smoke changes a newborn's DNA methylation in a set of locations across the genome, which forms an epigenetic signature reflecting prior exposure (Breton, Byun et al. 2009, Joubert, Haberg et al. 2012). This signature of prenatal exposure to maternal smoking persists through the first few years of childhood (Ladd-Acosta, Shu et al. 2016) and into adolescence (Lee, Richmond et al. 2015).

The conceptual framework presented below demonstrates three possible ways that infection, maternal immune activation, and brain and blood DNA methylation are related (Figure 4.1) (Ladd-Acosta 2015). Maternal immune activation is a useful construct to understand the detected associations between infections, fever, and immune markers such as CRP and autism. MIA may reflect a common pathway by which different prenatal inflammatory insults converge to affect the developing fetus. In model 1, MIA is responsible for epigenetic alterations in the developing brain, which are themselves causally related to or necessary for ASD development. It is most likely that the disease tissue of interest in the case of ASD is brain, not blood. MIA may additionally be responsible for epigenetic changes detectable in whole blood, but these may not be reflective of any changes present simultaneously in brain. If this is the case, changes in DNAm in the blood would be useful as a biomarker of MIA exposure, but would not necessarily themselves illuminate ASD pathogenesis.

In model 2, blood DNAm may reflect aspects of brain DNAm; differentially methylated regions generated in brain as a result of MIA exposure may spill over into

blood and be detectable (Davies, Volta et al. 2012). If this is the case, the magnitude of difference observed in blood between unexposed and exposed children is likely to be less than that present in the brain, but would still shed light on relevant epigenetic regions for understanding ASD pathology.

In model 3, neither brain nor blood DNAm are in the causal pathway of ASD development, but merely serve as biomarkers of an exposure that is causally related. Additionally, while not represented in the conceptual framework, is it possible that epigenetic changes are a consequence of ASD, and thus not in the causal pathway; they still could be associated with MIA exposure, however, if MIA is related to ASD risk. This possibility is difficult to assess without repeated DNAm measures at different ages.

In summary, DNA methylation differences in whole blood are most likely to serve as biomarkers of an exposure associated with autism spectrum disorder. They may also serve as proxies for the disease tissue of interest, if epigenetic changes in brain to some degree influence peripheral blood epigenetics. Differentiating between these two possibilities requires further investigation of a differentially methylated site with functional biological assays or comparison in different tissue types.

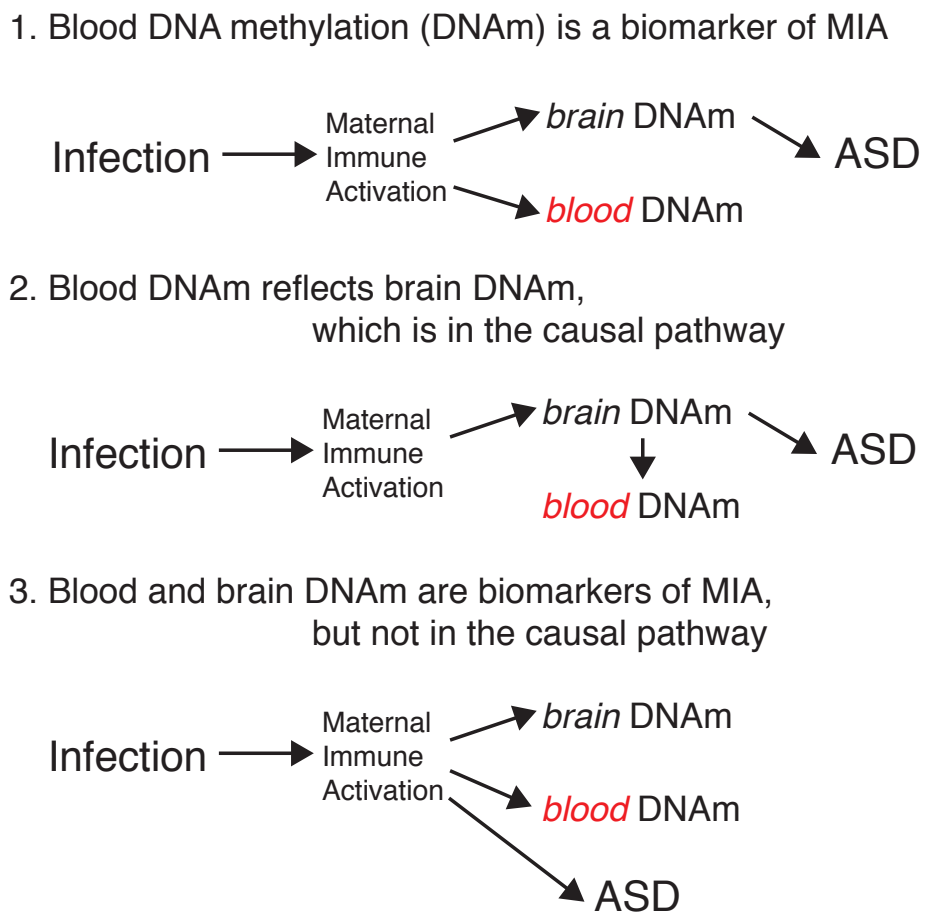


Figure 4.1: Three models for the potential relationship between infection, MIA, brain and blood DNA methylation, and Autism Spectrum Disorder.

4.2 Methods

4.2.1 Study to Explore Early Development

Johns Hopkins and Kennedy Krieger comprise the Maryland site of the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) in the Study to Explore Early Development (SEED), one of six national SEED sites (Figure 4.2). SEED is a national, CDC-funded case-control study with thorough phenotyping, retrospective data on prenatal exposures along with medical records, coupled with genotyping and epigenotyping arrays. The first phase of SEED (SEED I) collected data from over 2800 families (Schendel, Diguiseppi et al. 2012). Children were eligible for the ASD case group if they were born in a catchment area between 9/1/2003 and 8/31/2006 and utilized relevant ASD services such as early intervention, special education, hospitals, and clinics. ASD cases were thus 2-5 years old at the time of recruitment. General population controls from the same area and born in the same period were randomly sampled from state vital records. Potential cases and controls were invited to participate in the study and consenting families were enrolled, followed by clinical ASD diagnoses and extensive biological and epidemiological data collection.

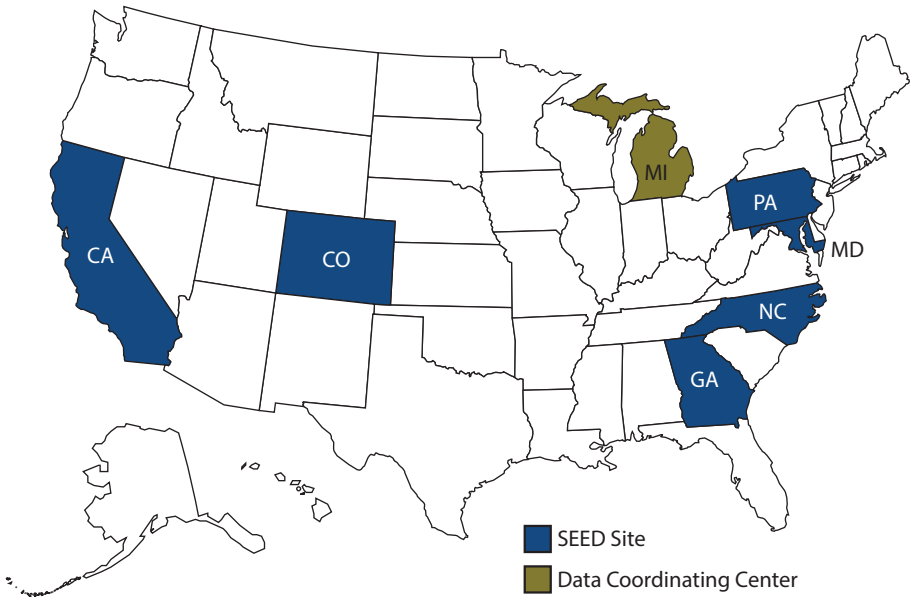


Figure 4.2: Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) and Data Coordinating Center for the Study to Explore Early Development (SEED), phase one.

4.2.2 Exposure assessment

We modeled prenatal exposure to maternal immune activation as exposure to infection during the perinatal period, according to maternal self-report provided via structured telephone interview at the time of enrollment in SEED. Thus exposure was assessed retrospectively, when the children were aged 2-5 years.

The SEED maternal exposure interview specifically asked about history of 36 different infections at any time during pregnancy and during each trimester. The most frequent categories of specific infections were genitourinary infection (bacterial vaginosis, candidiasis, chlamydia, genital herpes, HPV, pyelonephritis, trichomoniasis,

UTI) and respiratory infection (influenza, pneumonia, URI). There was also an opportunity for open-ended response for "other conditions," which were matched to appropriate ICD codes by the SEED Data Coordinating Center. ICD codes for infections were then extracted from these "free text" responses.

History of an infection was asked for the three months prior to conception (preconception or T0), during the first trimester (T1), during the second trimester (T2), during the third trimester (T3), or while breastfeeding. A dichotomous "yes/no" variable was generated for each time period based on maternal report of any kind of infection (viral, bacterial, or fungal; any organ system). Trimester-specific exposure was not mutually exclusive; a mother could report exposure for multiple trimesters. A dichotomous variable for "any infection at any time" during pregnancy was created based on the T1, T2, T3 variables.

4.2.3 Autism outcome

Autism outcome was assessed in SEED participants in a multistep process described in detail in prior literature (Wiggins, Reynolds et al. 2015). First, potential study participants were contacted by phone. During this invitation phone call, the family was asked to complete a brief ASD screener, the Social Communication Questionnaire (SCQ; Rutter, Bailey et al. 2003) (also see **Chapter 3** of this dissertation). A score equal to or greater than 11 or a prior ASD diagnosis was used to identify children for comprehensive developmental and ASD-specific evaluation. Children with no prior ASD diagnosis and an SCQ below 11 were asked to complete a shorter study visit with

cognitive, social, and motor developmental testing. A study algorithm based on prior literature, best practice guidelines, and clinical experience was developed to identify four groups of children: (1) ASD, (2) suspected ASD but incomplete study evaluation, (3) developmental disability or DD, and (4) population control.

We generated a binary variable based on these research classifications of ASD status. Children were given a '1' if they were identified as an ASD case or a suspected ASD case, and a '0' if they were a population control. No children in the DD group were selected for DNA methylation measurement, and are not included in the current analysis.

4.2.4 Epigenetic outcome data

Genomic DNA from whole blood was isolated using the QIAsymphony midi kit (Qiagen) at the Johns Hopkins Biological Repository (SEED Biorepository) as specified by the manufacturer. We bisulfite treated 500 ng of gDNA using the EZ DNA methylation kit Zymo Research Corp, Orange, CA, USA) according to the manufacturer's recommendations, as specified for downstream processing with 450K. To obtain genome-scale methylation measurements, bisulfite treated DNA samples were processed on the Infinium HumanMethylation450 BeadChip (Illumina, San Diego, CA, USA) at the Johns Hopkins SNP Center, in accordance with the manufacturer's recommendations, to obtain methylation data at 485,512 loci. Samples were randomized across and within plates to minimize potential batch and confounding effects. Sample replicates were included across plates for quality control to ensure methylation measurement reproducibility.

Quality control measures for the SEED 450k data were applied at both the sample and locus (probe) levels. As is standard in genetic and epigenetic analyses, these measures included removing sex-discordant, duplicate, or poorly performing samples; and removing poorly performing probes.

4.2.5 Statistical analyses

All statistical analyses were performed using R-3.1.x and 3.2.x and Bioconductor 3.0, specifically the packages *minfi* and *Bumphunter*, which are designed for the handling and analysis of epigenetic data obtained from the 450k platform (Gentleman, Carey et al. 2004, Jaffe, Murakami et al. 2012, Aryee, Jaffe et al. 2014, Huber, Carey et al. 2015).

Main effect analysis

Epigenetic data, ASD case status, and available covariate data were linked using a unique study family ID. The subset of children with epigenetic data, ASD case status, and all covariates was used. The association between child risk of ASD and maternal infection exposure prior to pregnancy, during each trimester or at any time during pregnancy, or while breastfeeding was assessed using unadjusted and adjusted logistic regression. Models were adjusted for child sex, study site (California, Colorado, Georgia, Maryland, North Carolina, and Pennsylvania), and the child's age in months at the time of their clinic visit. The threshold for significance for each association was a p value <0.05 and an odds ratio 95% confidence interval excluding 1.

Epigenetic data preprocessing

Raw intensity data (.idat files) were imported using the read.450k.exp() function in *minfi*. Beta values were obtained by dividing the methylated probe intensity by the overall probe intensity (the sum of the methylated and unmethylated probe intensities); beta values represent percent methylation, from 0 to 100% methylated. Beta values were then logit-transformed to obtain M-values, which are normally distributed and more appropriate for our statistical methods (Du, Zhang et al. 2010). Data were then normalized with noob, which uses background probes and dye-bias correction to normalize data based on technical variation (Triche, Weisenberger et al. 2013).

Batch correction

Because a second batch of SEED samples was run several years after the first batch, statistical control for the unmeasured confounders associated with different run dates had to be performed. We used two different strategies for batch correction: ComBat (Chen, Grennan et al. 2011) and surrogate variable analysis with the R package *sva* (Leek, Johnson et al. 2012). While ComBat directly adjusts for the known batches, *sva* will detect latent sources of variation in a study sample; the latent variation can then be visualized against known sources of technical variability, including batch. These two methods of batch correction were performed in parallel, and the downstream effects on single-site association methods and genomic control were compared.

This epigenome-wide association study used both single site association methods as well as regional methods (Rakyan, Down et al. 2011).

Single site association analysis

Differentially methylated positions (DMPs) were identified using linear models, both unadjusted and adjusted for appropriate confounders, as accounted for by the surrogate variables estimated by *sva*. Potential confounders of the relationship between maternal infection and offspring methylation include technical variation, such as that due to batch or sample position on the 450k array, as well biological variation unrelated to our question of interest, including cell type composition, sex, and ancestry.

Epigenetic analyses conducted in whole blood are complicated by the heterogenous cell population in the sample. We generally wish to ensure that any detected methylation differences are not driven by a change in the relative proportion of the cells comprising the sample. There are two strategies to address this: (1) cell types in a sample are measured with a laboratory assay and then compared across exposed and unexposed individuals, to ensure that cell type proportions are stable despite exposure status, (2) cell type composition is estimated based on a subset of probes in the 450k platform that have been validated to predict relative cell proportions (Jaffe and Irizarry 2014). If not adjusted for, differences in cell type proportions among exposed and unexposed samples could result in a spurious association between exposure and methylation. Because we were interested in epigenetic differences that are not caused by differences in the frequency of one particular cell type over another, we ensured that the estimated surrogate variables adjusted for cell type.

Similarly, because we were interested in epigenetic differences related to infection exposure and not the individual's sex, we also ensured that the surrogate variables accounted for sex.

Additionally, if a study is not homogenous with regard to ancestry, it is important to assess for the possibility of confounding by population stratification. We found that our estimated surrogate variables did not account for ancestry. However, in SEED, principal components of ancestry have been derived from genotyping data; we included the first three principal components of ancestry in a sensitivity analysis to ensure that our results were stable.

The model for the single site association analysis was structured so that methylation at a single probe site was the outcome, represented as a linear combination of variables representing the exposure of interest and the potential confounders:

$$\text{M-value} = B_0 + B_1\text{infection} + B_2x_2 + \ldots + B_nx_n$$

Using the lmFit() and eBayes() functions in *minfi*, an empirical Bayes method was used to obtained moderated t-statistics where the sample variances were moved towards a pooled value, allowing more stable inferences (Smyth 2004). Model performance was evaluated by assessing genomic inflation; appropriate control of genomic inflation (lambda ≤ 1.05) suggests that a valid model has been developed to fit the study data. Essentially, over 455,000 linear models were run estimating the association between prenatal exposure to infection at a particular time point and methylation at a given probe site. The p-values for the association between prenatal infection exposure and methylation at each probe site were extracted and DMPs ranked based on significance. P-

values were adjusted for multiple testing within the model, generating False Discovery Rates (FDR) or q-values using the Benjamini-Hochberg correction. We took a q-value < 0.05 to be evidence for significance, so we expect that out of the null hypotheses we reject, 5% will be false discoveries. Based on the number of tested probes and an alpha level of 0.05, we also calculated a Bonferroni significance threshold (1.0973e-07).

Regional association analysis

Next, we used a region-finding approach that takes advantage of probe clustering on the 450k platform to detect larger areas of consistent methylation differences between cases and controls (differentially methylated regions, DMRs) (Jaffe, Murakami et al. 2012). This is conducted with the bumphunter() function as implemented in the *minfi* package, and was performed with adjustment for the same set of surrogate variables used in the single-site association analysis. Parameters for bumphunting were set with a max distance between probes within a cluster of 250 bp; significance was estimated by family-wise error rates (FWER) for 1000 bootstraps. Empirical p-values and FWER were estimated for the first percentile of differentially methylated regions, on the basis of length + area under the curve as well as area alone. The properties of the empirical p-values are not well understood, and so we based our assessment of significance on the FWER. The family wise error rate is the proportion of null bootstraps for which any DMR has a value as or more extreme than the DMR of interest (see Appendix B for an example of an Epigenome Wide Association Study, utilizing both single site and regional analysis).

4.3 Results

4.2.1 Sample description

The frequencies of exposed and unexposed children in SEED for infection during pregnancy, trimester-specific infection, and before and after pregnancy (while breastfeeding) are shown in Table 4.1. The epigenetic analytic sample included 620 males and 309 females. Two subjects were missing data on ASD case status, but among the remaining subjects there were 500 population controls and 427 ASD cases. Five subjects were missing data on age, but the mean age among the remaining 924 subjects was 59.3 months (4.9 years) with a range of 34.6 to 70.7 months (2.9 to 5.9 years). The epigenetic sample reflects the age and sex distribution of the larger SEED sample (Wiggins, Levy et al. 2015). There were 171 subjects from the California SEED site; 200 from the Colorado site; 141 from the Georgia site; 163 from the Maryland site; 124 from the North Carolina site; 128 from the Pennsylvania site; and two subjects without an annotated study site.

Table 4.1: Prenatal exposures for children with epigenetic data in SEED

	No, unexposed	Yes, exposed
Infection in 3 months prior to conception	841 (90.5%)	88 (9.5%)
Infection at any time in pregnancy	589 (63.4%)	340 (36.6%)
Trimester 1	799 (86.0%)	130 (14.0%)
Trimester 2	754 (81.2%)	175 (18.8%)
Trimester 3	718 (77.3%)	211 (22.7%)
Infection while breastfeeding	815 (87.7%)	114 (12.3%)

4.2.2 Analysis of the main effect

The subset of children with epigenetic data, ASD case status, and all covariates was used (n=924). Two children with epigenetic data were missing information on ASD case status, and three children with epigenetic data were missing information on study site or age at clinic visit. In this subset of children with epigenetic data and infection exposure data, after adjustment for child sex, child age at study participation, and study site, we found a significant association between child ASD status and maternal infection in the three months prior to conception; and during trimester 1, 2, or 3 (Table 4.2).

Table 4.2: Unadjusted and adjusted OR for ASD risk after prenatal infection exposure with 95% confidence intervals

	Unadjusted (95% CI)	Adjusted[a] (95% CI)	exposed (n)	unexposed (n)
Infection in 3 months prior to conception	**1.78 (1.15 - 2.81)**	**1.78 (1.11 - 2.89)**	88	836
Infection at any time in pregnancy	1.31 (0.998 - 1.71)	1.31 (0.98 - 1.74)	584	340
Trimester 1	**1.60 (1.10 - 2.33)**	**1.64 (1.11- 2.46)**	794	130
Trimester 2	**1.63 (1.17 - 2.28)**	**1.61 (1.13 - 2.30)**	749	175
Trimester 3	**1.47 (1.08 - 2.01)**	**1.41 (1.02 - 1.96)**	713	211
Infection while breastfeeding	**1.52 (1.03 - 2.26)**	1.52 (0.998 - 2.34)	810	114

[a] adjusted for child sex, child age at clinic visit, and study site

Bold denotes statistically significant associations

4.2.3 Epigenetic data description

The 450k array was run on 1014 samples, including placenta and liver control samples and 12 cross-array duplicate samples. After data preprocessing and quality control measures, 970 samples and 455,664 probes remained for further analysis (Figure 4.3). To run a complete case analysis, our analytic sample was less the 41 individuals for whom we have no infection exposure information (final analytic sample size = 929). These data were collected in two batches—batch one (12/11/12 − 1/10/13; 574 samples)

and batch two (6/18/15 − 7/2/15; 355 samples)—and were balanced on infection exposure status (Table 4.3).

Figure 4.3

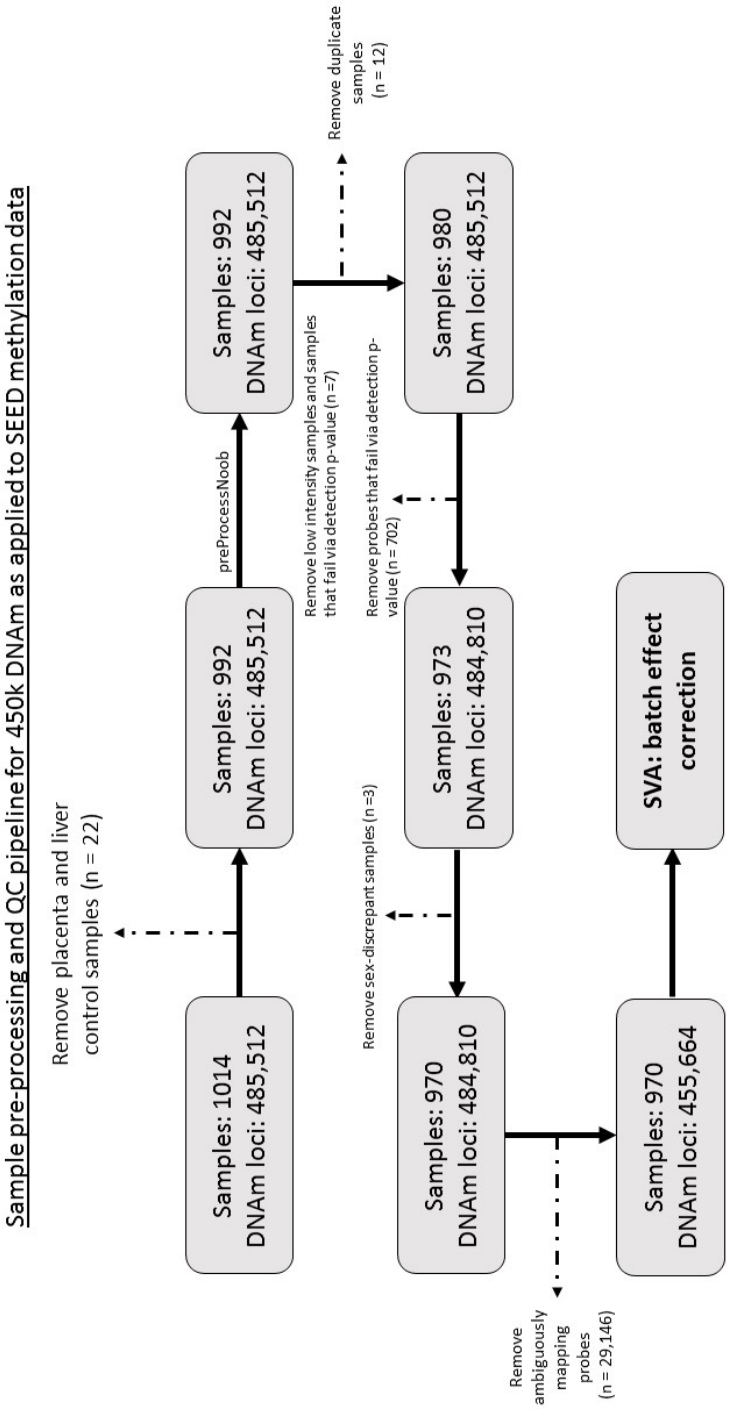


Table 4.3: Distribution of infection exposure by 450k batch

Batch	unexposed	exposed
1	363 (63.2%)	211 (36.8%)
2	226 (63.7%)	129 (36.3%)

4.2.4 Batch Correction

We noted significant p-value deflation in both completely unadjusted data (Figure 4.4) and data corrected for batch with ComBat (Figure 4.5; data shown for the analysis of any infection at any time during pregnancy, though the same pattern was seen for the other exposure variables). With a single-site association analysis, the expected versus observed p-values can be compared with a qq plot (Casella and Berger 2002). An estimate of the p-value inflation can be assessed with the lambda statistic; in genome-wide association studies, significant p-value inflation can be observed and the expectation is that adjustment for appropriate confounders will decrease the lambda < 1.05 to minimize false positive discovery. It was unexpected here to see deflation, rather than inflation. Correction for batch with ComBat and subsequent direct adjustment for confounders (including sample sex and estimated cell type composition) did not alleviate the p-value deflation.

We then pursued using *sva* (surrogate variable analysis) to estimate surrogate variables for latent sources of variation in the data, including batch. For exposure to any infection at any time during pregnancy, 30 surrogate variables were estimated. Then each surrogate variable was regressed on potential explanatory variables, including batch,

array plate, array row position, sample sex, and estimated cell types (CD8 T cells, CD4 T cells, natural killer cells, B cells, monocytes, and granulocytes) (Figure 4.6). The first 18 surrogate variables adequately captured technical sources of variation (plate, batch, row 6) and biological sources of variation different from the main question we were interested in (sex, cell type composition) (Figure 4.7). We also found that there was a significant association between three of the surrogate variables and ASD case status, indicating that including the surrogate variables in our analysis also accounted for variation in methylation that is due to the ASD case-control design of SEED (Figure 4.8).

We did note that the surrogate variables were not associated with ancestry (see Figure 4.8, demonstrating that the surrogate variables estimated for third trimester infection exposure are not significantly associated with ancestry principal components derived from SNP array data). In a sensitivity analysis, we adjusted for the first three ancestry principal components in addition to the first 18 surrogate variables; for third trimester infection exposure, the two differentially methylated were loci were still the top ranked probes, and the first still had a q-value < 0.05. Accounting for ancestry did not substantially alter our results.

Moving forward, we adjusted for the first 18 surrogate variables when testing the association between infection during or prior to pregnancy and DNA methylation in offspring at age 2-5 years. This method of adjustment improved the p-value deflation observed in the completely unadjusted data and in the data corrected for batch with ComBat (Figure 4.9 cf. Figure 4.4 and Figure 4.5; Table 4.4).

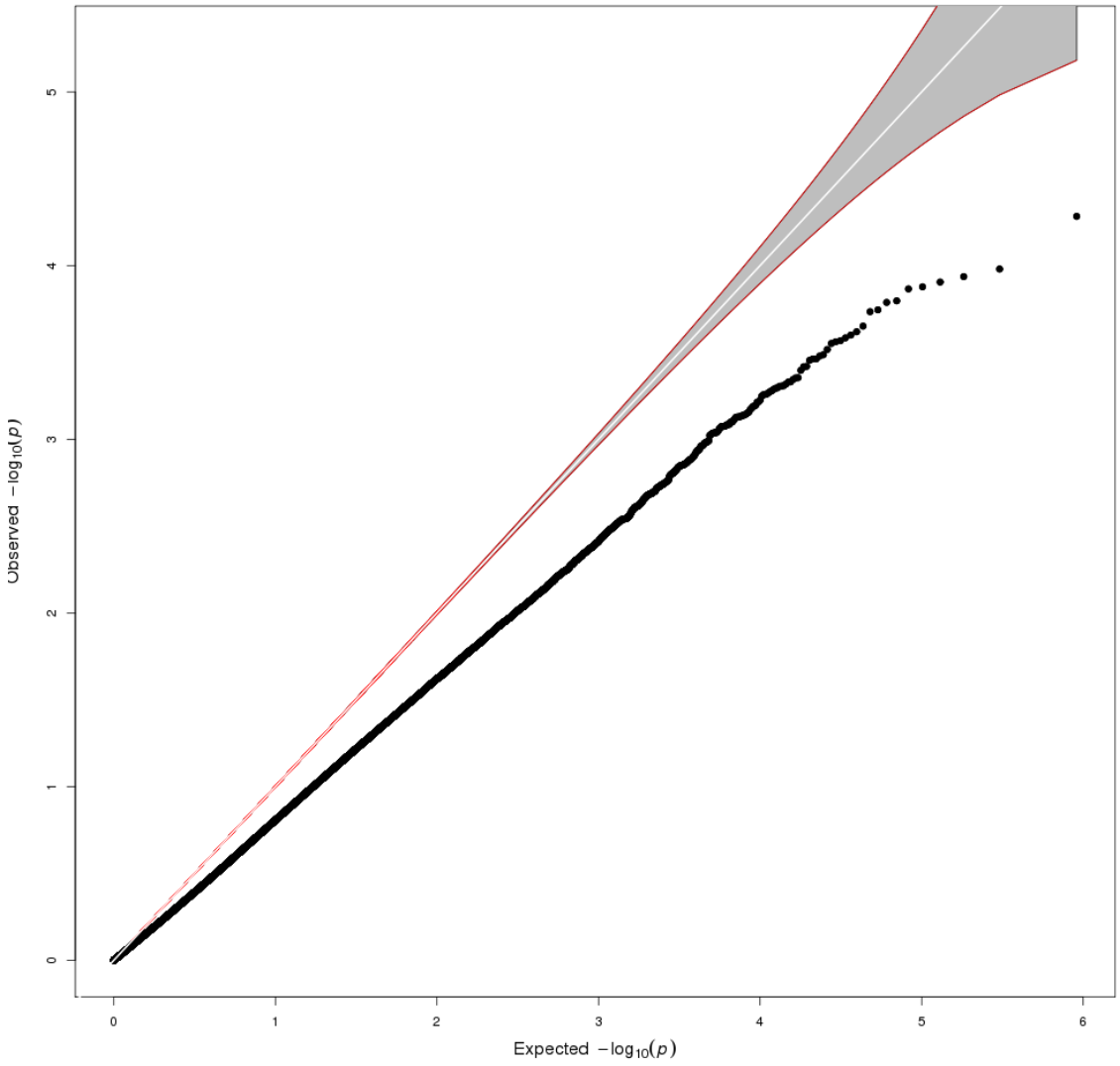


Figure 4.4: qq plot for the single site association analysis of any infection at any time during pregnancy, without batch correction or adjustment for confounders. Lambda is estimated to be 0.75, which quantifies the decrease in observed -log10(p-values) compared to the expectation.

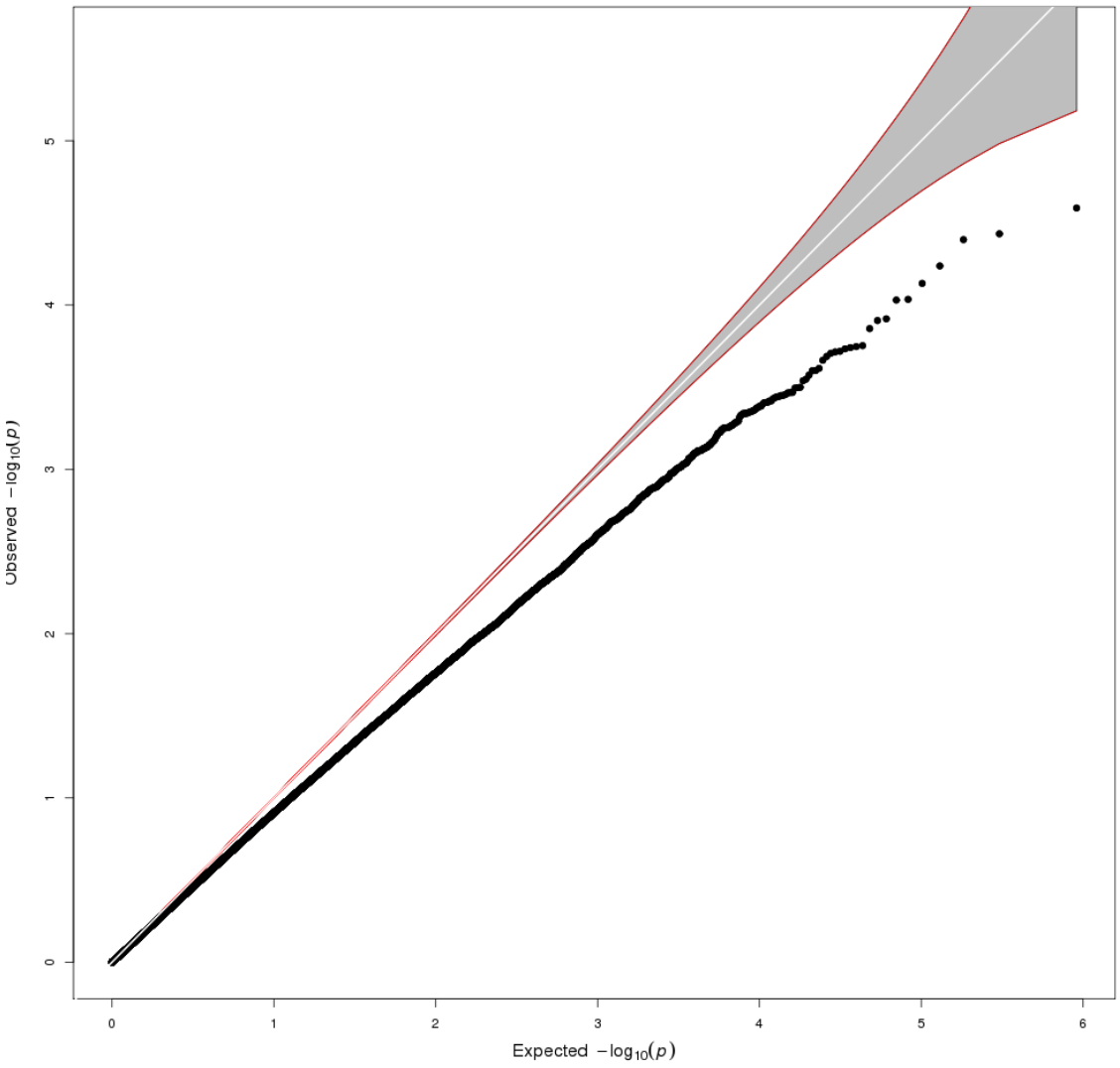


Figure 4.5: qq plot for the single site association analysis of any infection at any time during pregnancy, with batch correction performed via ComBat but without adjustment for other confounders. Lambda is estimated to be 0.86, which quantifies the decrease in observed -log10(p-values) compared to the expectation.

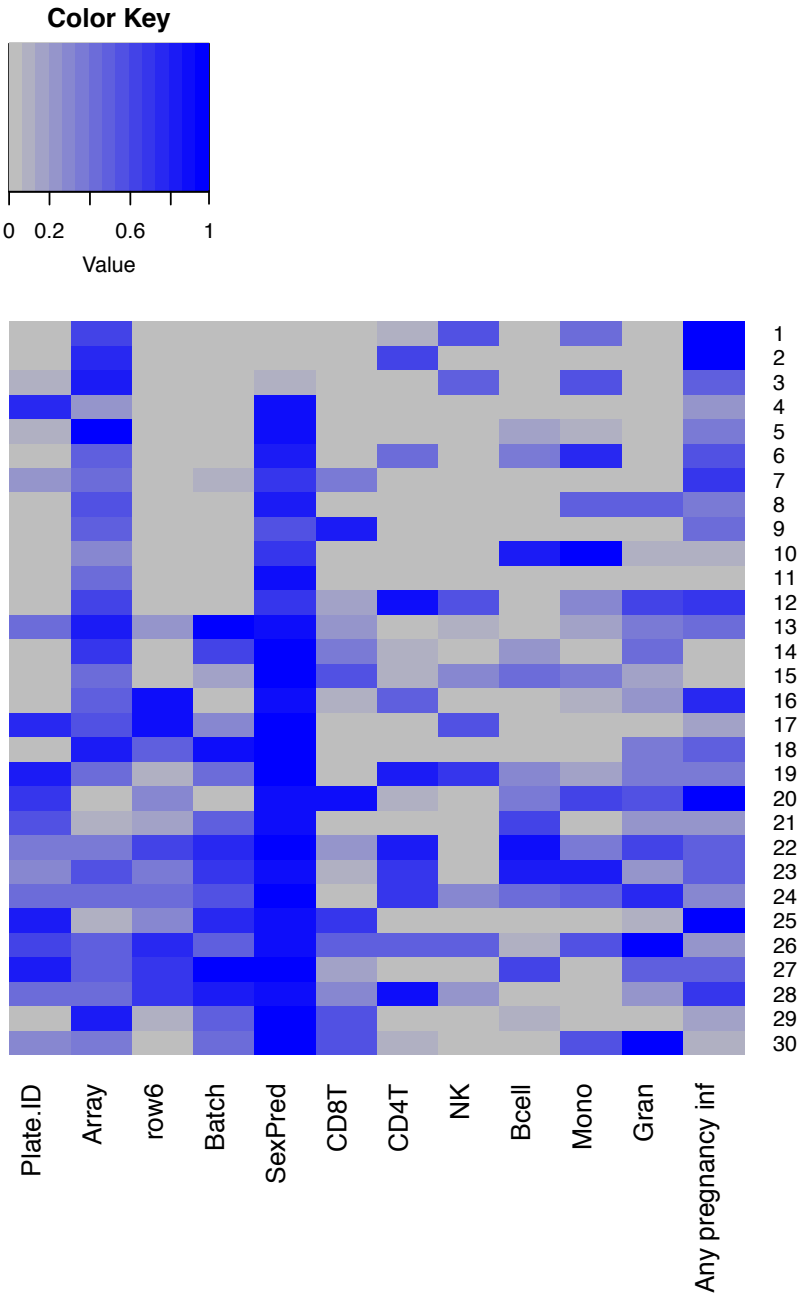


Figure 4.6: Heat map showing the degree of significance of the association between a surrogate variable (1-30, vertical axis) estimated for the comparison of subjects exposed and unexposed to any infection at any time during pregnancy, and an explanatory variable (horizontal axis). Blue shading increases as the p-value for the association increases (association moves farther from the significance threshold).

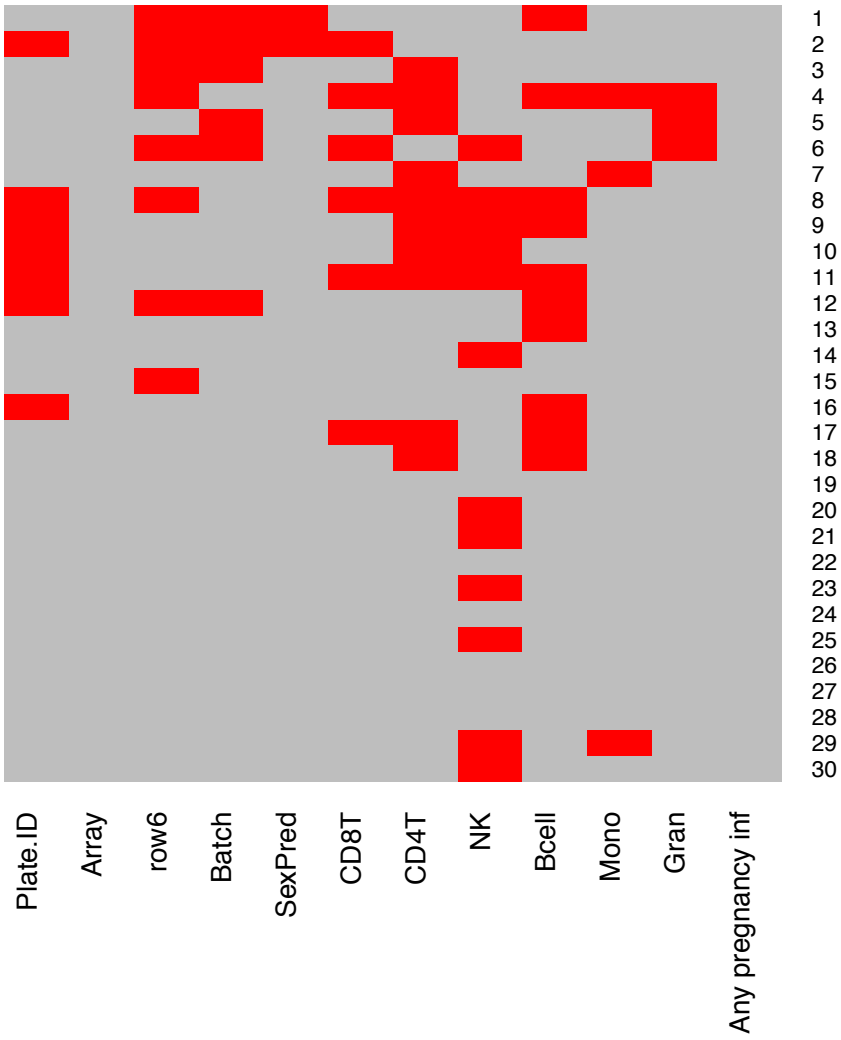


Figure 4.7: Associations that meet the Bonferroni threshold for significance are marked in red (p-value < 0.000139; corrected for testing 30 surrogate variables against 12 explanatory variables, or 360 tests). Note, no surrogate variables are significantly associated with variable "array," which represents the specific sample well, or "Any pregnancy inf" (infection, any time during pregnancy), which is the source of biological variation we set out to protect in estimating the latent sources of technical and unwanted biological variation. The first 18 surrogate variables adequately capture variation contributed by plate, row 6, batch, sex, and estimated cell type composition.

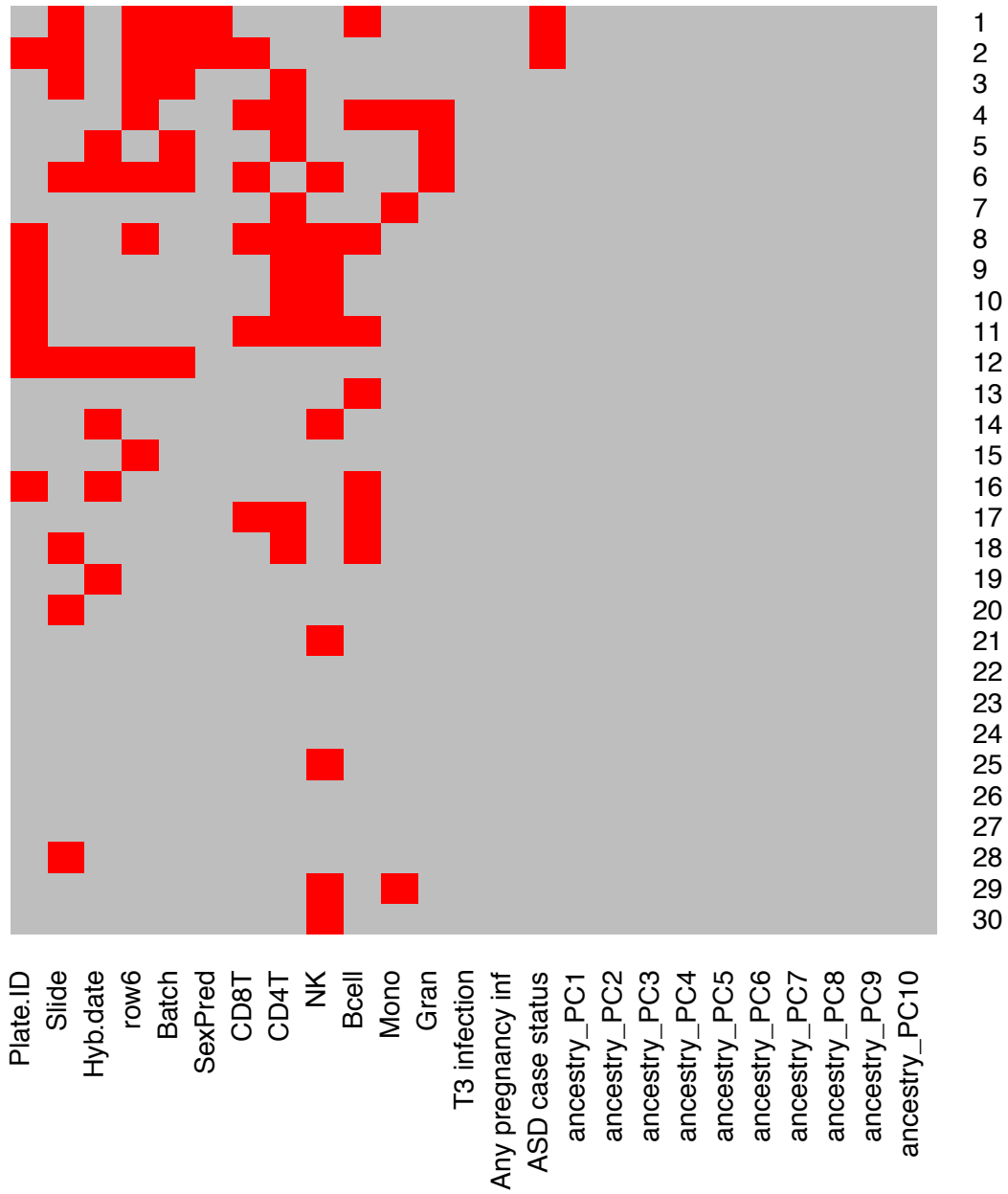


Figure 4.8: Associations that meet the Bonferroni threshold for significance are marked in red. These surrogate variables were estimated for the comparison of third trimester infection exposed and unexposed children. The surrogate variables do account for variation due to the ASD case status, but do not account for ancestry (ancestry_PC1-PC10).

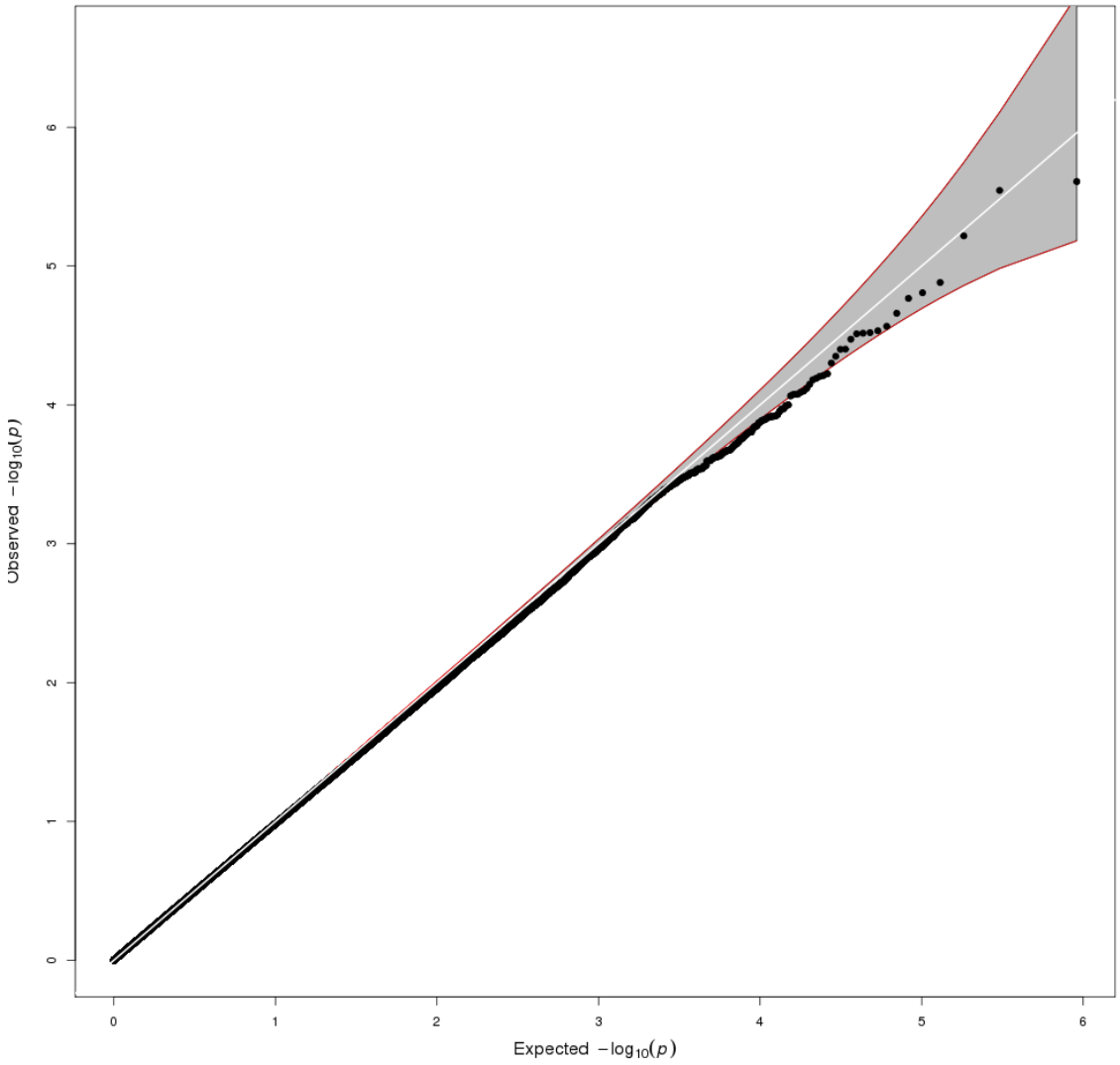


Figure 4.9: qq plot for the single site association analysis of any infection at any time during pregnancy, with batch correction and adjustment for other confounders performed via *sva*. Lambda is estimated to be 0.98.

Table 4.4: Lambda values for the single site analysis for each exposure variable

	Unadjusted	Adjusted for first 18 surrogate variables
Infection in 3 months prior to conception	0.76	1.10
Infection at any time in pregnancy	0.75	0.98
Trimester 1	0.70	0.99
Trimester 2	0.76	1.00
Trimester 3	0.84	1.06
Infection while breastfeeding	0.61	1.01

4.2.5 Single-site association analysis (Differentially Methylated Positions)

After adjustment for the first 18 surrogate variables, we estimated the association between methylation at a particular locus and each exposure variable. For children exposed to any infection at any time during gestation, we did not find any differentially methylated positions (Table 4.5). For children whose mothers reported an infection in the three months prior to their conception, we found a single differentially methylated probe (Table 4.6). This was located at genomic position chr5:172903876. For children exposed to any infection during trimester 1 or trimester 2, we did not find any differentially methylated positions (Table 4.7, Table 4.8). For children exposed to any infection during trimester 3, we found two differentially methylated positions (Table 4.9): chr3:12947823

within the gene body of *IQSEC1*, and chr1:110306507 within the gene body of *EPS8LS*. For children whose mothers reported an infection while breastfeeding, we did not find any differentially methylated positions; however, the top ranked probe (q-value < 0.1) is also within the gene body of *IQSEC1*, though upstream of the probe identified in the trimester 3 infection exposure analysis (Table 4.10). We performed a BLAT analysis of the hybridization sequence of the three significantly differentially methylated probes, and found that the full sequence unambiguously mapped to the region expected based on the Illumina probe annotation (obtained using the getAnnotation() function in the *minfi* package) (Table 4.11).

Table 4.5: Top ranked 450k probes for any infection at any time during pregnancy

Chromosome	Position	Gene	ΔM[a]	*P* value	Adjusted q-value	Annotated SNP[b]
chr17	3828268	ATP2A3	-0.43	2.46E-06	0.64977	
chr1	181683239	CACNA1E	0.62	2.85E-06	0.64977	
chr4	137732125		-0.37	6.05E-06	0.91944	
chr8	47120700		0.33	1.31E-05	0.99996	
chr8	145579317	FBXL6; C8ORFK29	-0.12	1.56E-05	0.99996	
chr8	35685207		1.26	1.71E-05	0.99996	
chr2	158175831	ERMN	-0.14	2.19E-05	0.99996	
chr16	81040735	CENPN	-0.12	2.71E-05	0.99996	
chr8	2585976		-3.00	2.92E-05	0.99996	
chr14	24458099	DHRS4L2	0.54	3.01E-05	0.99996	

[a] Difference in mean percent methylation levels between patients exposed and unexposed to infection during gestation. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals exposed to infection compared to those unexposed.

[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the *minfi* Bioconductor package.

Table 4.6: Top ranked DMPs after maternal infection exposure 3 months prior to conception

Chromosome	Position	Gene	ΔM[a]	*P* value	Adjusted q-value	Annotated SNP[b]
chr5	**172903876**		**-2.68**	**1.15E-08**	**0.005**	
chr12	54385625	MIR196A2	-1.95	3.05E-07	0.069	
chr14	89715336	FOXN3	-3.01	1.14E-06	0.161	rs115466766
chr20	36767986	TGM2	-0.48	1.67E-06	0.161	
chr20	981673	RSPO4	-3.69	1.77E-06	0.161	rs11906926
chr1	21504121	EIF4G3	-2.12	2.25E-06	0.171	
chr20	34287060	ROMO1; NFS1	1.99	4.70E-06	0.238	
chr6	28864188		1.39	5.22E-06	0.238	
chr15	25312161	SNORD116-7; SNORD116-5	-1.42	5.43E-06	0.238	
chr7	6268584	CYTH3	1.14	6.05E-06	0.238	

[a] Difference in mean percent methylation levels between patients exposed and unexposed to infection during gestation. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals exposed to infection compared to those unexposed.

[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the *minfi* Bioconductor package.

Table 4.7: Top ranked DMPs after maternal infection exposure during trimester 1

Chromosome	Position	Gene	ΔM[a]	*P* value	Adjusted q-value	Annotated SNP[b]
chr8	101225344	SPAG1	2.00	2.91E-06	0.65434	
chr1	145415441	HFE2	0.76	6.03E-06	0.65434	
chr8	101225252	SPAG1	2.14	7.43E-06	0.65434	
chr5	86709126	CCNH	0.88	8.07E-06	0.65434	
chr22	19436867	C22orf39	0.50	1.09E-05	0.65434	
chr6	42018203	TAF8	0.59	1.09E-05	0.65434	
chr7	134001865	SLC35B4	0.14	1.11E-05	0.65434	
chr7	155534636	RBM33	-0.19	1.31E-05	0.65434	
chr4	184021351	WWC2; C4orf38	-0.53	1.41E-05	0.65434	
chr8	72987762	TRPA1	0.28	1.61E-05	0.65434	

[a] Difference in mean percent methylation levels between patients exposed and unexposed to infection during gestation. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals exposed to infection compared to those unexposed.

[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the *minfi* Bioconductor package.

Table 4.8: Top ranked DMPs after maternal infection exposure during trimester 2

Chromosome	Position	Gene	ΔM[a]	*P* value	Adjusted q-value	Annotated SNP[b]
chr14	90240124	HCG26	-0.29	1.42E-07	0.06473	
chr6	31438151	C21orf63	-2.18	1.80E-06	0.41114	rs9267136
chr21	33784769	MDP1	-0.04	6.12E-06	0.85501	
chr14	24685281	GABBR1	-0.08	1.42E-05	0.85501	
chr6	29578496		0.28	1.62E-05	0.85501	
chr11	64270338	WNT16	-0.73	1.62E-05	0.85501	
chr7	120969079	MSH4	0.69	2.11E-05	0.85501	
chr1	76262373	EPN1	-0.70	2.14E-05	0.85501	
chr19	56196747	MRPL12	0.03	2.15E-05	0.85501	
chr17	79670933	HCG26	-0.14	2.32E-05	0.85501	

[a] Difference in mean percent methylation levels between patients exposed and unexposed to infection during gestation. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals exposed to infection compared to those unexposed.

[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the *minfi* Bioconductor package.

Table 4.9: Top ranked DMPs after maternal infection exposure during trimester 3

Chromosome	Position	Gene	ΔM[a]	*P* value	Adjusted q-value	Annotated SNP[b]
chr3	**12947823**	**IQSEC1**	**-0.66**	**3.02E-08**	**0.01378**	
chr1	**110306507**	**EPS8L3**	**-0.87**	**1.59E-07**	**0.03633**	
chr19	51220202	SHANK1	-0.15	5.12E-07	0.06654	
chr16	84746995	USP10	-0.54	6.50E-07	0.06654	
chr11	130013355	APLP2	0.31	7.81E-07	0.06654	
chr17	38468610	RARA	-0.66	8.76E-07	0.06654	
chr8	37758453	RAB11FIP1	1.22	2.78E-06	0.18120	
chr17	40175841	NKIRAS2	1.02	3.47E-06	0.19754	
chr2	10182525	KLF11	-1.09	4.63E-06	0.23236	
chr17	62252524	TEX2	2.00	5.56E-06	0.23236	

[a] Difference in mean percent methylation levels between patients exposed and unexposed to infection during gestation. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals exposed to infection compared to those unexposed.

[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the *minfi* Bioconductor package.

Table 4.10: Top ranked DMPs after maternal infection exposure while breastfeeding

Chromosome	Position	Gene	ΔM[a]	*P* value	Adjusted q-value	Annotated SNP[b]
chr3	13063165	IQSEC1	1.01	2.44E-07	0.08581	
chr6	6971315		-0.31	3.77E-07	0.08581	
chr6	28831393		0.62	2.83E-06	0.33016	
chr10	104210692	C10orf95	0.30	2.90E-06	0.33016	
chr12	8219358	C3AR1	2.41	4.71E-06	0.34278	
chr7	2644645	IQCE	-0.43	6.12E-06	0.34278	
chr13	73356092	PIBF1;DIS3	0.52	6.78E-06	0.34278	
chr6	33384537	CUTA	1.04	6.91E-06	0.34278	
chr13	43395572		-1.21	7.04E-06	0.34278	
chr8	117778752	UTP23	0.11	8.06E-06	0.34278	

[a] Difference in mean percent methylation levels between patients exposed and unexposed to infection during gestation. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals exposed to infection compared to those unexposed.

[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the *minfi* Bioconductor package.

Table 4.11: BLAT (BLAST-like alignment tool) analysis

Illumina probe name	Illumina manifest Annotated Chr:position	exposure period	sequence	BLAT output
cg01802043	chr5:172903876	T0	CGTGTTTTCATGTCCA GTGTCTAGCAGGGTGA ATGGTACACAGTAGGC AT	100.0% identity to 50 bp seq on chr 5
cg01460382	chr3:12947823	T3	GAAAGGTCAGATGTGC TCTGGACCAGATGGGG GCTGCAAGCTCCCCAG CG	100.0% identity to 50 bp seq on chr 3 100.0% identity to 20 bp seq on chr 3 100.0% identity to 20 bp seq on chr 17
cg00515905	chr1:110306507	T3	GGCCGGGTGCCTGGTC CCCCCAGGAGGCTGGT CTTGGAGCAGGTGGTC CG	100.0% identity to 50 bp seq on chr 1

For the three probes that were differentially methylated based on infection exposure, we plotted the percent methylation for four groups: ASD cases who were exposed; ASD cases who were unexposed; population controls who were exposed; and population controls who were unexposed. For the probe that was differentially methylated based on maternal preconception infection, cases and controls were separated based on their preconception (T0) exposure status (Figure 4.10). There were 52 children with ASD who were T0 exposed; 375 children with ASD who were T0 unexposed; 36 population controls who were T0 exposed; and 464 population controls who were T0 unexposed. For the two probes that were differentially methylated based on trimester 3 (T3) exposure, we separated ASD cases and population controls based on their T3 exposure status (Figure 4.11 and Figure 4.12). There were 113 children with ASD who

were T3 exposed; 314 children with ASD who were T3 unexposed; 98 population controls who were T3 exposed; and 402 population controls who were T3 unexposed.

Next, we subset the data to those children whose mothers reported never having an infection during their pregnancy (n=589) and to those children whose mothers reported an infection during trimesters 1, 2, and 3 (n=59) and visualized the methylation at each of the three positions identified as significant in the full sample (Figures 4.13 - 4.15).

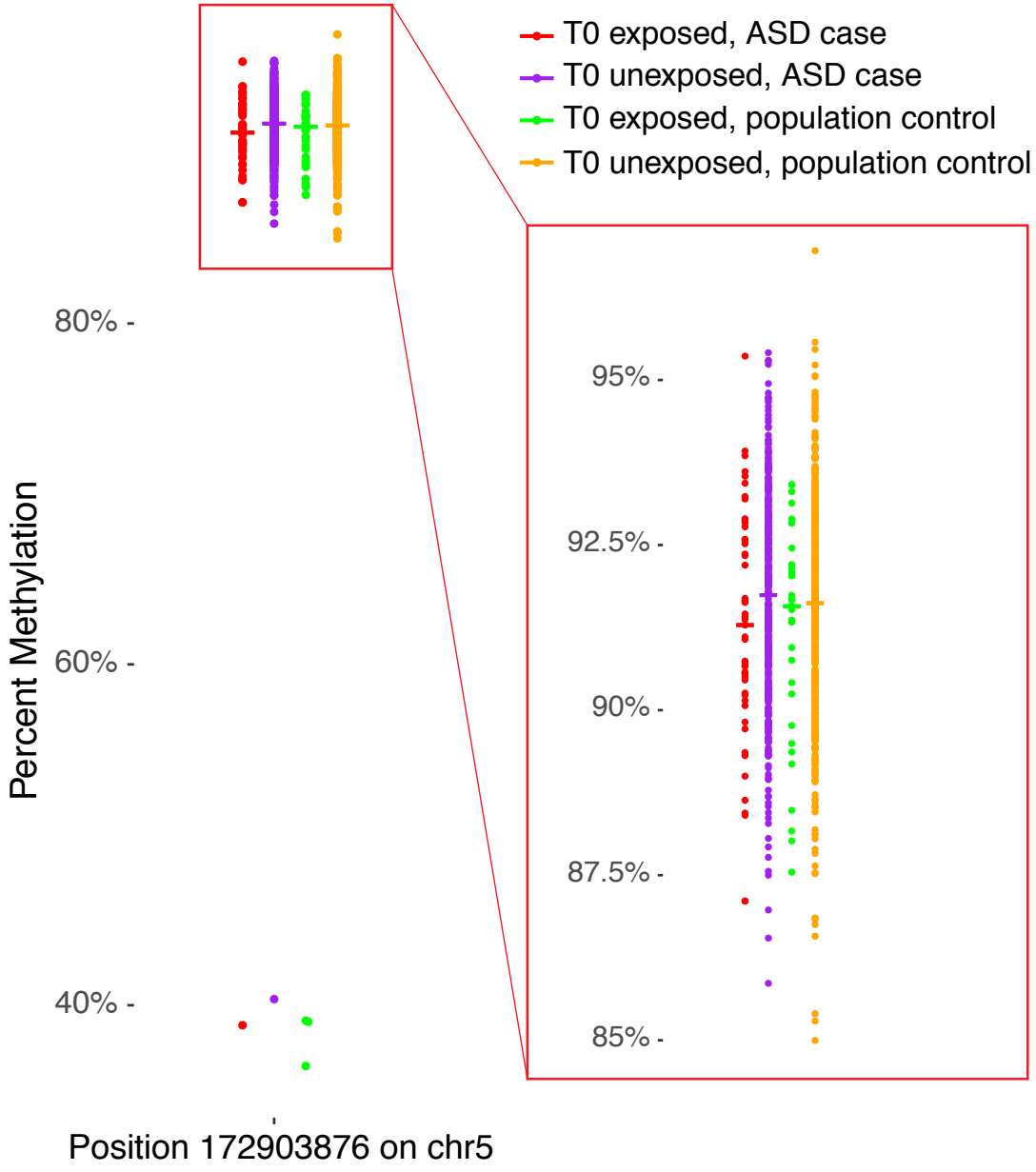


Figure 4.10: DNA methylation of children age 2-5 enrolled in SEED on chromosome 5 at position 172903876, as measured by a probe on the 450k array. 927 children are plotted based on maternal report of preconception infection and ASD case status. Median percent methylation for each exposure-ASD group is marked by a short horizontal line. Five samples have methylation <50%.

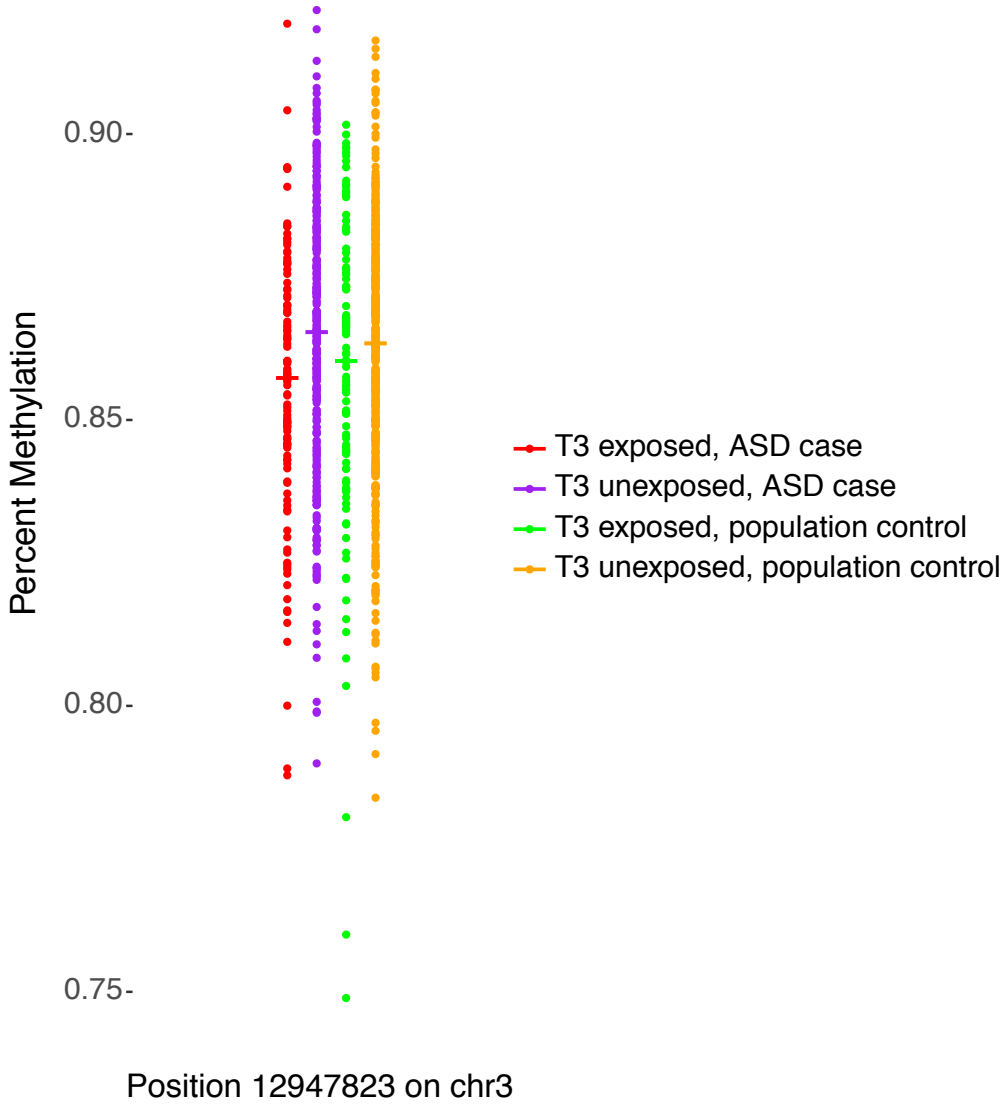


Figure 4.11: DNA methylation of children age 2-5 enrolled in SEED on chromosome 3 at position 12947823, as measured by a probe on the 450k array. 927 children are plotted based on T3 exposure and ASD case status.

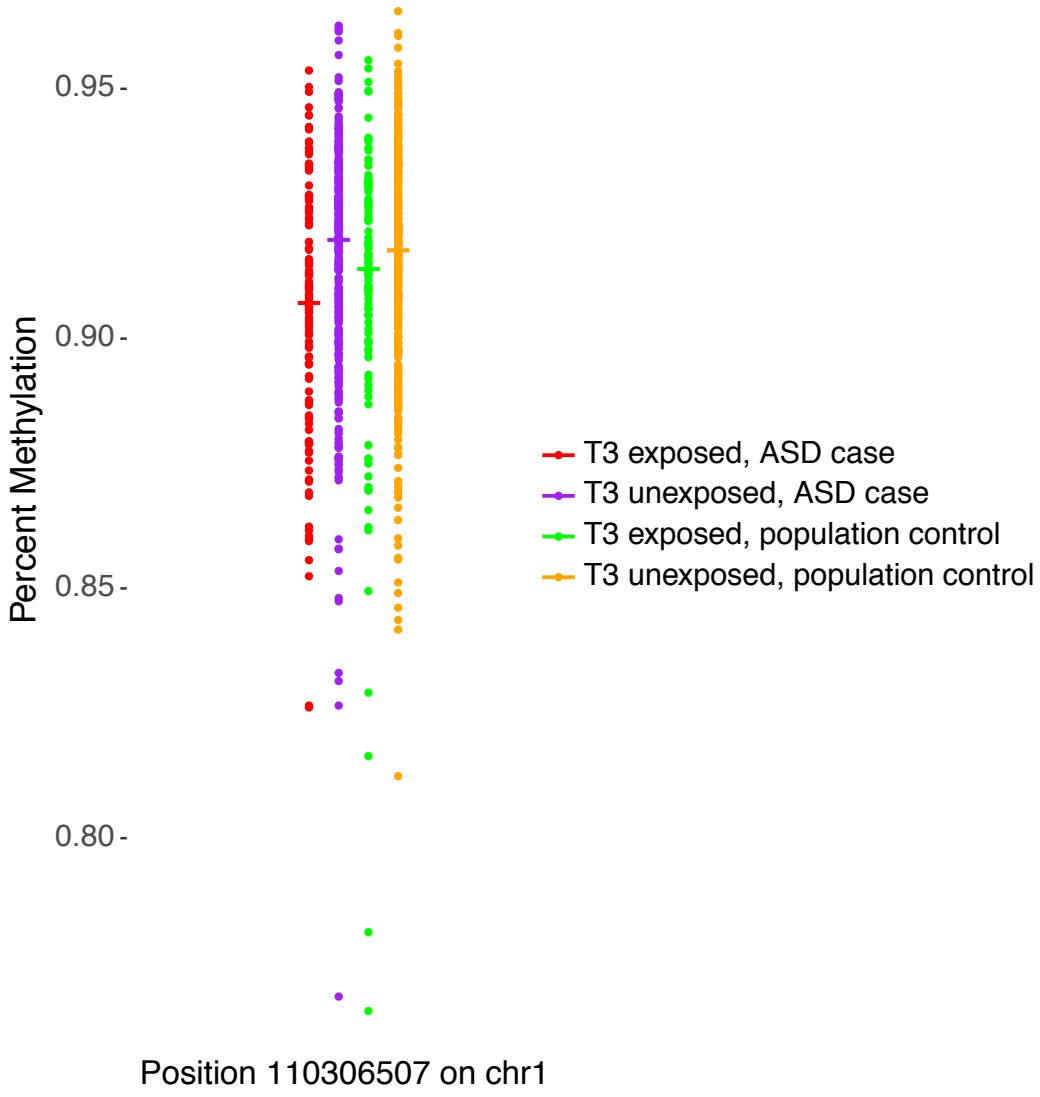


Figure 4.12: DNA methylation on chromosome 1 at position 110306507 as measured by a probe on the 450k array. 927 children are plotted based on T3 exposure and ASD case status.

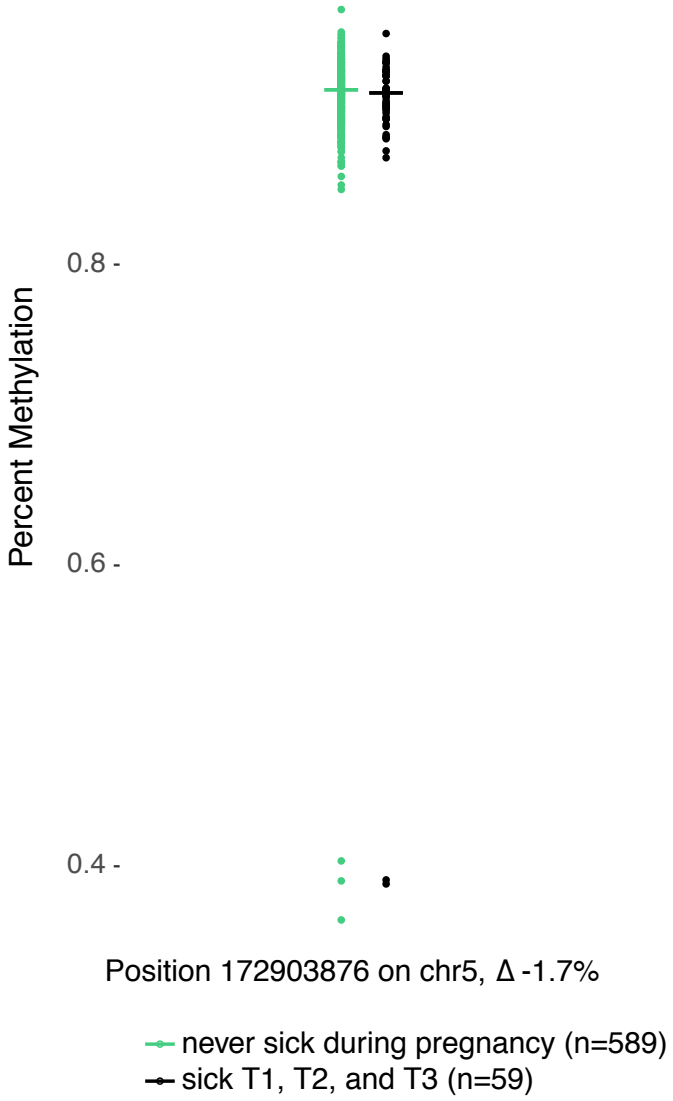


Figure 4.13: DNA methylation on chromosome 5 at position 172903876, among children whose mothers were never sick during their pregnancy (n=589) and children whose mothers reported being sick every trimester of their pregnancy (n=59). On average, children whose mothers were sick throughout pregnancy have 1.7% less methylation at this locus.

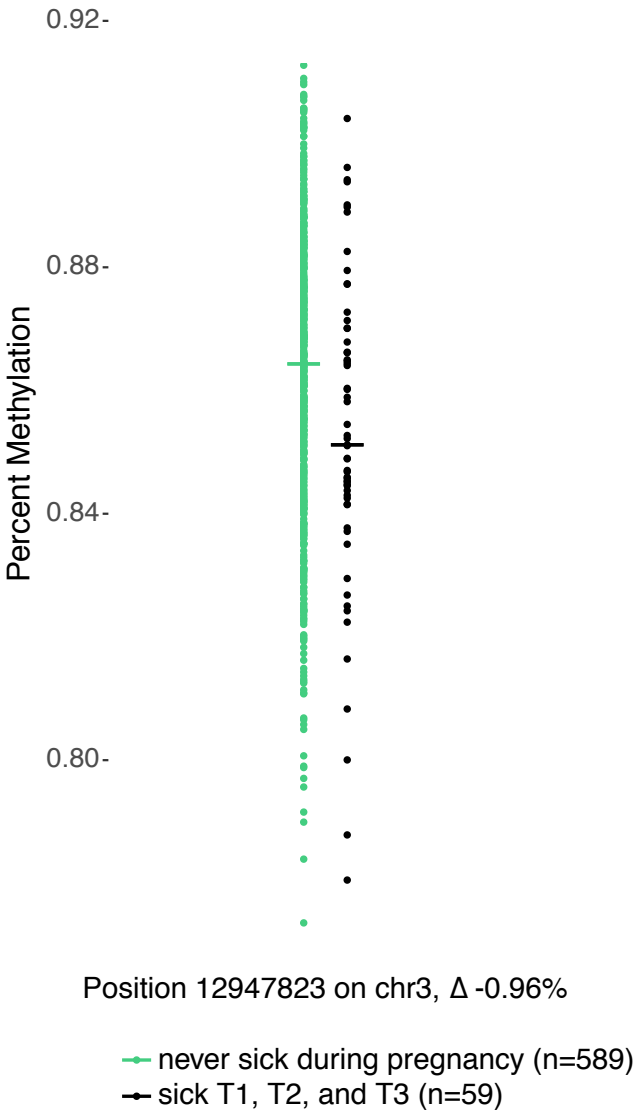


Figure 4.14: DNA methylation at chr3:12947823, among children whose mothers were never sick during their pregnancy (n=589) and children whose mothers reported being sick every trimester of their pregnancy (n=59). On average, children whose mothers were sick throughout pregnancy have 0.96% less methylation at this locus.

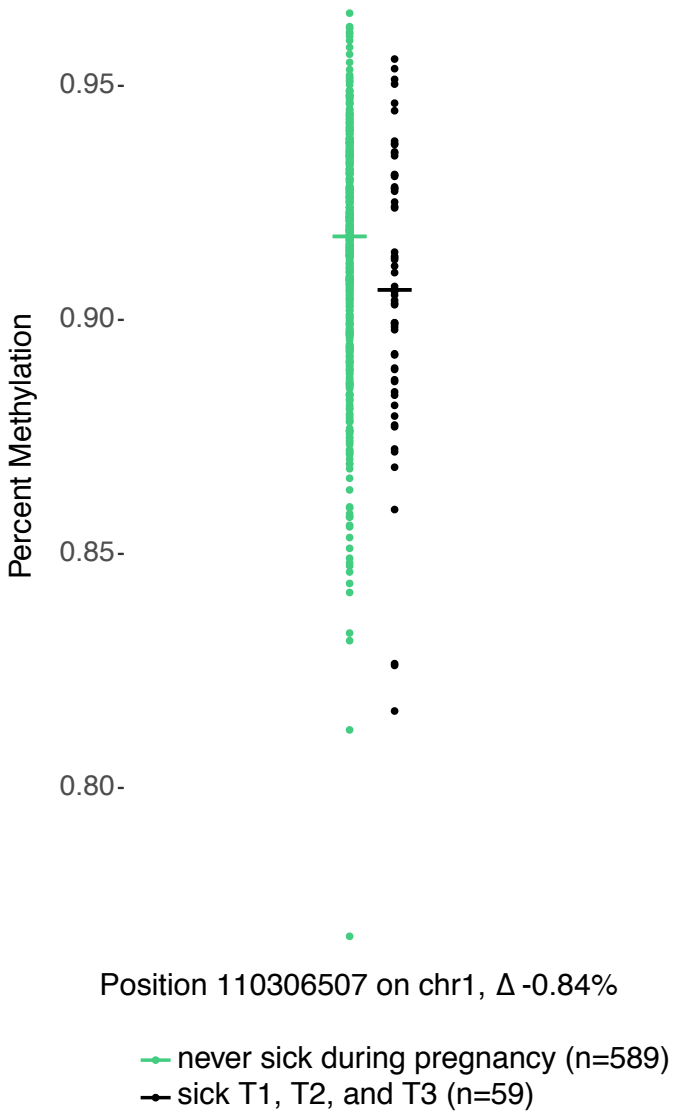


Figure 4.15: DNA methylation at chr1:110306507, among children whose mothers were never sick during their pregnancy (n=589) and children whose mothers reported being sick every trimester of their pregnancy (n=59). On average, children whose mothers were sick throughout pregnancy have 0.96% less methylation at this locus.

4.2.6 Regional analysis (Differentially Methylated Regions)

Next, we used a region finding approach for our six exposure variables (Tables 4.12-4.17). While none of the identified regions are statistically significantly different between exposure groups, there are several regions that were identified in the top ranked DMRs across multiple comparisons: a 57 bp region in the promoter of *SDHAP3* (Figure 4.16), a 775 bp region at the 5' end of *RUFY1* (Figure 4.17), and a 120 bp region overlapping an exon of *PIEZO1* (Figure 4.18).

Table 4.12: Top ranked DMRs for any infection at any time during pregnancy

Chromosome	Position	Distance[a]	Gene[b]	Location	FWER	FWER area
chr21	43989949	ss	SLC37A1	inside intron	0.16700	1
chrX	70712215: 70712810	595	TAF1	inside intron	0.57800	0.855
chr7	22481962	ss	STEAP1B	inside intron	0.58300	1
chr6	41068553: 41068752	199	NFYA	inside exon	0.75100	0.962
chr3	27674461: 27674472	11	EOMES	downstream	0.79000	1
chr19	57742112: 57742444	332	AURKC	overlaps 5'	0.80400	0.987
chr5	178986131: 178986906	775	RUFY1	overlaps 5'	0.84200	0.999
chr5	1594676: 1594733	57	SDHAP3	promoter	0.89100	1
chr9	115771599	ss	ZNF883	inside intron	0.89400	1
chr12	131118426: 131118654	228	RIMBP2	inside intron	0.95100	1

[a] Number of base pairs that the DMR covers. 'ss' indicates that the DMR was composed of a single probe site, rather than a cluster of probes.

[b] Nearest annotated gene.

Table 4.13: Top ranked DMRs for any infection prior to conception

Chromosome	Position	Distance[a]	Gene[b]	Location	FWER	FWER area
chr5	1594676:1594733	57	SDHAP3	promoter	0.30800	0.954
chr1	75198211:75199117	906	CRYZ	overlaps 5'	0.31300	0.559
chr14	24779793:24780734	941	LTB4R2	overlaps exon upstream	0.44800	0.75
chr2	54086854:54087343	489	ASB3	overlaps 5'	0.49900	0.795
chr12	7781004:7781431	427	APOBEC1	downstream	0.59700	0.984
chr3	182817190:182817584	394	MCCC1	overlaps 5'	0.68700	0.945
chr16	88803931:88804051	120	PIEZO1	overlaps exon downstream	0.83600	1
chr11	5617812:5617926	114	TRIM6-TRIM34	overlaps 5'	0.83900	1
chr15	100821466:100821466	ss	ADAMTS17	inside exon	0.85300	1
chr20	46415320:46415320	ss	SULF2	inside exon	0.89300	1

[a] Number of base pairs that the DMR covers. 'ss' indicates that the DMR was composed of a single probe site, rather than a cluster of probes.

[b] Nearest annotated gene.

Table 4.14: Top ranked DMRs for any infection exposure during the first trimester

Chromosome	Position	Distance[a]	Gene[b]	Location	FWER	FWER area
chr6	33049983: 33050124	141	HLA-DPB1	inside intron	0.20800	0.999
chr19	13875014: 13875137	123	MRI1	promoter	0.41300	0.999
chr5	135415948: 135416613	665	VTRNA2-1	covers	0.41800	0.698
chr1	75198211: 75199117	906	CRYZ	overlaps 5'	0.43100	0.715
chr6	30039374: 30039524	150	RNF39	overlaps exon upstream	0.45900	0.751
chr14	24779793: 24780734	941	LTB4R2	overlaps exon upstream	0.62000	0.903
chr8	101224915: 101225361	446	SPAG1	overlaps exon upstream	0.62300	0.966
chr6	29648590: 29649092	502	ZFP57	upstream	0.69300	0.966
chr11	124613956	ss	NRGN	inside intron	0.83800	1
chr6	28945182: 28945507	325	ZNF311	downstream	0.91500	0.994

[a] Number of base pairs that the DMR covers. 'ss' indicates that the DMR was composed of a single probe site, rather than a cluster of probes.

[b] Nearest annotated gene.

Table 4.15: Top ranked DMRs for any infection during the second trimester

Chromosome	Position	Distance[a]	Gene[b]	Location	FWER	FWER area
chr6	32064212:32064660	448	TNXB	inside exon	0.31700	0.627
chr10	49654342	ss	ARHGAP22	inside exon	0.63700	1
chr3	195489708:195490309	601	MUC4	overlaps exon downstream	0.64100	0.884
chrX	70712215:70712810	595	TAF1	inside intron	0.73800	0.933
chr6	30039027:30039206	179	RNF39	inside exon	0.81500	0.957
chr8	10049871	ss	MSRA	inside intron	0.85700	1
chr17	5403053:5403516	463	LOC728392	overlaps two exons	0.89600	1
chr1	240620177	ss	FMN2	inside intron	0.90200	1
chr11	66317822	ss	ACTN3	inside intron	0.90200	1
chr8	143751796:143751801	5	PSCA	inside exon	0.93700	1

[a] Number of base pairs that the DMR covers. 'ss' indicates that the DMR was composed of a single probe site, rather than a cluster of probes.

[b] Nearest annotated gene.

Table 4.16: Top ranked DMRs for any infection during the third trimester

Chromosome	Position	Distance	Gene[a]	Location	FWER	FWER area
chr6	30039376: 30039476	100	RNF39	overlaps exon upstream	0.42800	0.726
chr16	88803931: 88804051	120	PIEZO1	overlaps exon downstream	0.48600	1
chr6	29648590: 29649092	502	ZFP57	upstream	0.49000	0.767
chr8	143751796: 143751801	5	PSCA	inside exon	0.50700	1
chr6	32551749: 32552453	704	HLA-DRB6	overlaps 5'	0.53300	0.842
chr5	178986131: 178986906	775	RUFY1	overlaps 5'	0.61500	0.877
chr6	29648379: 29648525	146	ZFP57	upstream	0.72500	0.955
chr5	1594676: 1594733	57	SDHAP3	promoter	0.74100	0.995
chr1	110254709: 110254896	187	GSTM5	overlaps 5'	0.75200	0.991
chr1	248100345: 248100614	269	OR2L13	overlaps 5'	0.75600	0.961

[a] Nearest annotated gene.

Table 4.17: Top 10 ranked DMRs for any infection while breastfeeding

Chromosome	Position	Distance[a]	Gene[b]	Location	FWER	FWER area
chr8	1140574	ss	ERICH1-AS1	downstream	0.13000	1
chr1	248100228: 248100614	386	OR2L13	overlaps 5'	0.13400	0.473
chr1	205818956: 205819609	653	PM20D1	overlaps 5'	0.15400	0.532
chr15	28200653	ss	OCA2	inside intron	0.53100	1
chr2	113992762: 113993313	551	PAX8	covers exon(s)	0.63100	0.884
chr1	146549909: 146549940	31	NBPF13P	downstream	0.73000	1
chr4	165878037: 165878219	182	TRIM61	inside intron	0.77300	0.988
chr17	57053: 57120	67	RPH3AL	downstream	0.86400	1
chr17	6558064: 6558440	376	MIR4520-1	close to 3'	0.94200	1
chr5	1594676: 1594733	57	SDHAP3	promoter	0.94900	1

[a] Number of base pairs that the DMR covers. 'ss' indicates that the DMR was composed of a single probe site, rather than a cluster of probes.

[b] Nearest annotated gene.

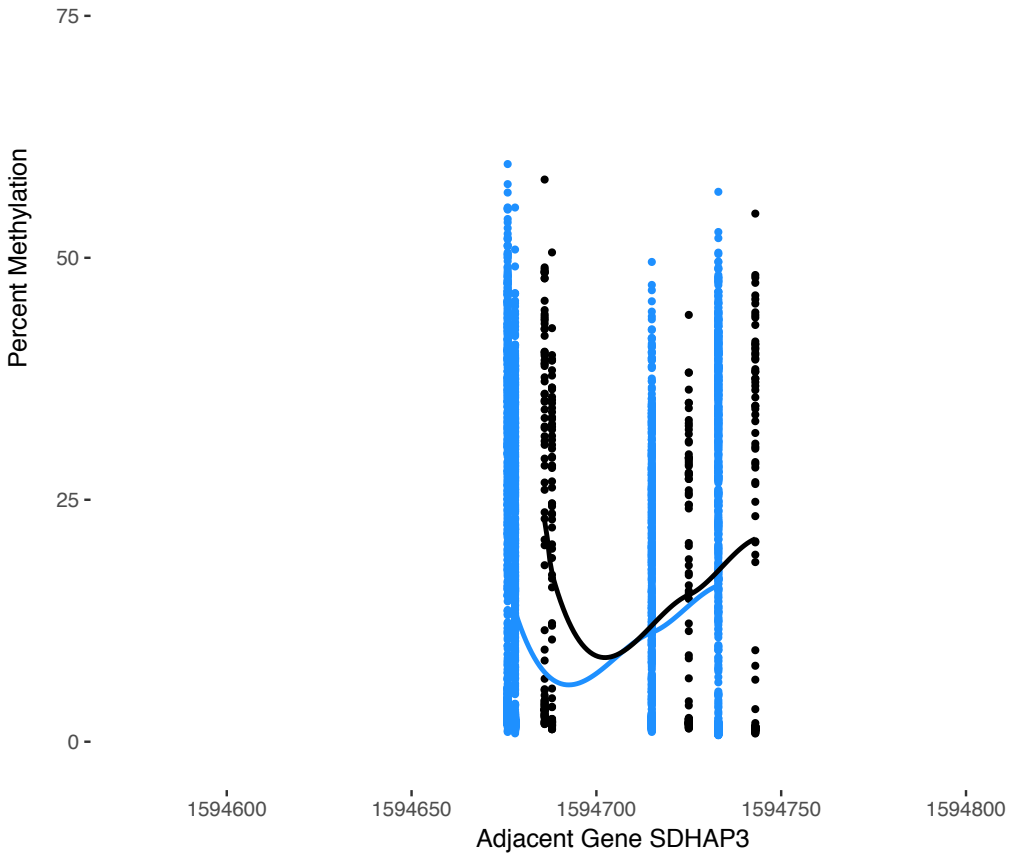


Figure 4.16: A 57 bp region in the promoter of *SDHAP3* that was in the top 10 ranked regions for the comparison of any infection during pregnancy, preconception infection, infection during T3, and infection while breastfeeding, though FWER > 0.1 for all comparisons. Plotted is the comparison of preconception maternal infection exposed (black) and unexposed (blue).

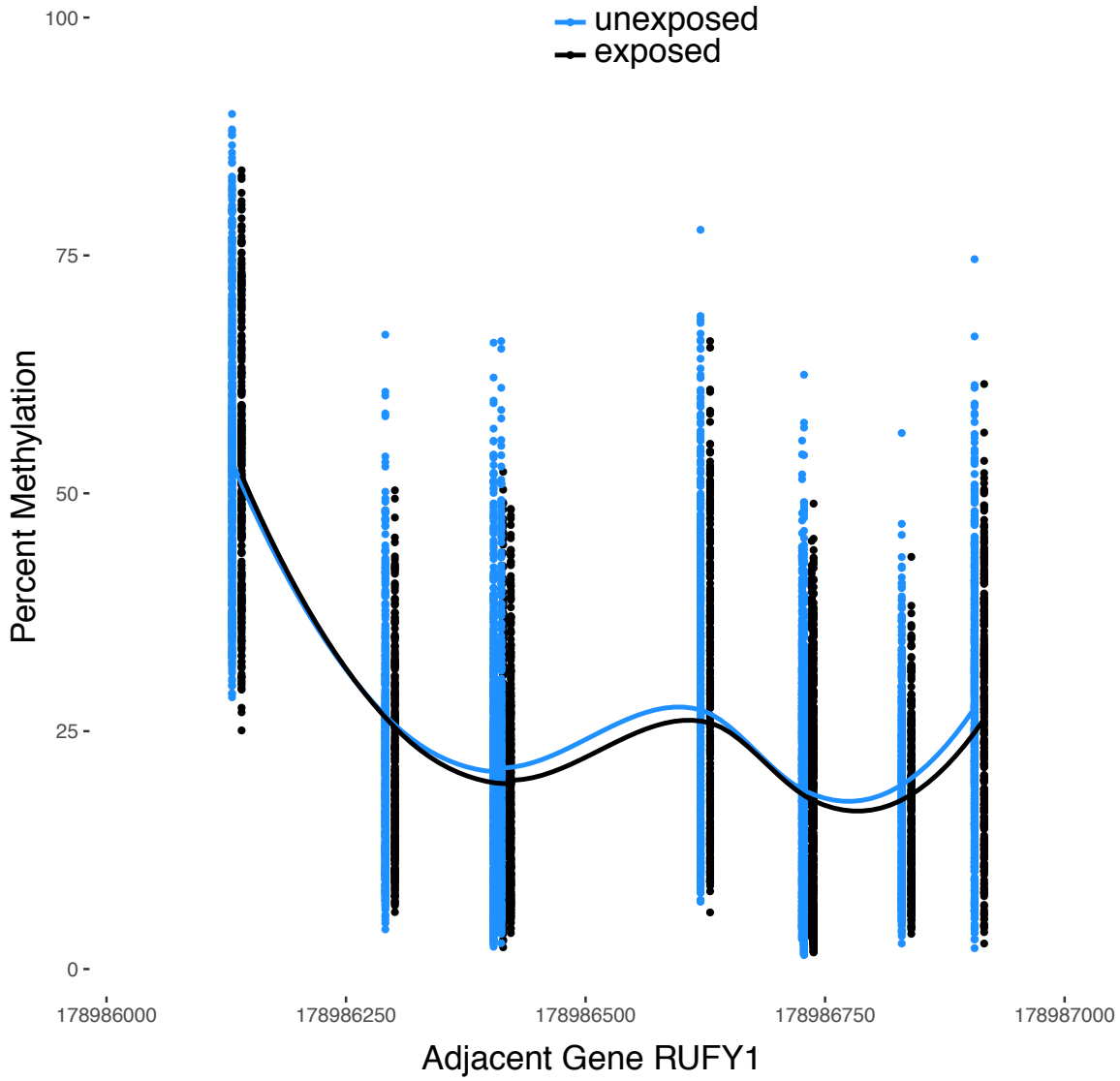


Figure 4.17: A 775 bp region at the 5' end of *RUFY1* that was in the top ranked regions for the comparison of any infection at any time during pregnancy and infection during T3, though FWER > 0.1 for all comparisons. Plotted is the comparison of third trimester infection exposed (black) and unexposed (blue).

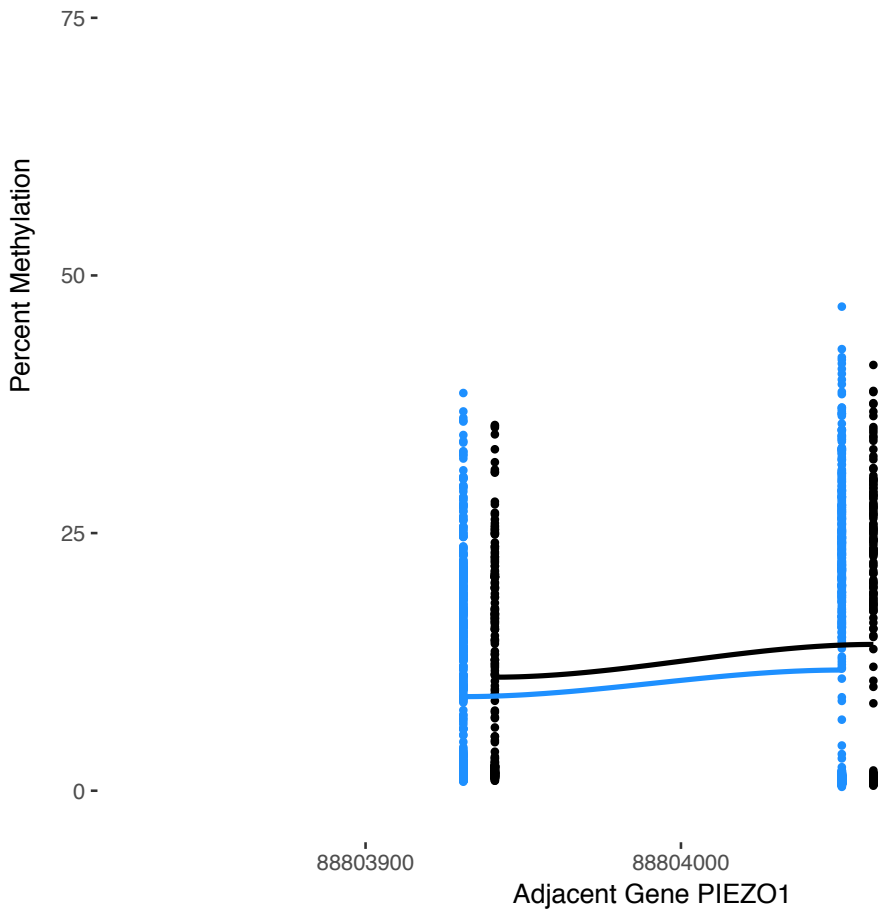


Figure 4.18: A 120 bp region overlapping an exon of *PIEZO1* that was in the top 10 ranked regions for the comparison of any infection prior to conception and infection during the third trimester, though FWER > 0.1 for all comparisons. Plotted is the comparison of third trimester exposed (black) and unexposed (blue).

4.4 Discussion

We tested for associations between prenatal infection exposure and DNA methylation in childhood blood in an agnostic, genome-scale approach. Such an epigenome-wide approach is superior to a candidate region study, as it tests for associations we may not expect based on our current imperfect understanding of the science. We represented maternal immune activation as infection during pregnancy for two reasons: (1) infection is likely to be the most common cause of MIA, as its reported prevalence is up to 60% in both retrospective and prospective interviews of pregnant women (Collier, Rasmussen et al. 2009), and (2) it is a prenatal exposure that is commonly collected in epidemiological studies, allowing for comparison with the literature.

We found no probes that were differentially methylated due to exposure to any infection at any time during pregnancy; exposure to any infection during the first or second trimester; or exposure to any infection when breastfed. We also found no regions that had a family wise error rate less than 0.05 for differential methylation in areas of probe clustering, for any of the six exposures of interest.

However, we found a single differentially methylated probe (q-value < 0.05, p-value < 1.0973e-07), among children whose mothers reported an infection in the three months prior to their conception. This was located at position chr5:172903876, which is in an open sea intergenic region that was included on the 450k array due to the presence of an enhancer site. The ENCODE project predicts that this position is near an enhancer-

like region in human astrocytes, a non-neuronal cell type found predominantly in synapses in the brain and spinal cord that may be involved in autism etiology (Consortium 2012, Blanco-Suarez, Caldwell et al. 2017). It is possible that this exposure is a proxy for general maternal health prior to conception, rather than specific to infection; it is also possible that report of an infection prior to conception actually overlaps with very early pregnancy.

We also found two probes that were differentially methylated in children who were exposed to an infection during the third trimester of their gestation. The first is at position chr3:12947823 (q-value < 0.05, p-value < 1.0973e-07), which is within the gene body of *IQSEC1*. *IQSEC1* is thought to be involved in synaptic transmission, as both a scaffolding and signaling protein (Um 2016). The second is at position chr1:110306507 (q-value < 0.05), within the gene body of *EPS8LS*; this gene is not well studied but is likely involved in actin regulation, and the region is predicted to be an enhancer in muscle and skin tissues and hematopoietic stem cells (Offenhauser, Borgonovo et al. 2004, Consortium 2012). While the absolute difference in percent methylation was modest when comparing exposed and unexposed in the full sample, the differences were more marked when looking specifically at children whose mothers reported infections throughout pregnancy compared to those whose mothers had no infections. This may reflect a dose-response relationship between a cumulative prenatal infection exposure and methylation at the identified sites. Additionally, small absolute differences in DNA methylation may still have profound effects on RNA expression. They could also represent a significant change in a rare cell population that nonetheless has important

biological functions (for example, a change in T_H17 cells). Small absolute changes in DNA methylation in response to prenatal environmental exposures may be the expected response; DNA methylation perturbations on the order of magnitude encountered in cancer research may not be compatible with life in a developing fetus (Breton, Marsit et al. 2004).

DNA methylation was measured in whole blood, which may not be an appropriate indicator of DNA methylation in brain tissue across the genome (Hannon, Lunnon et al. 2015, Bakulski, Halladay et al. 2016). However, a number of probe sites on the 450k platform show strong correlation between blood and brain regions, including prefrontal cortex, across individuals; while the absolute degree of methylation may or may not be comparable between blood and brain, inter-individual differences in methylation are conserved across tissues. While methylation in whole blood at positions chr3:12947823 and chr1:110306507 is not tightly correlated with brain DNA methylation in matched samples, it is at position chr5:172903876 (Hannon, Lunnon et al. 2015) (Table 4.18 and Figure 4.19; see the Blood Brain DNA Methylation Comparison Tool at http://epigenetics.essex.ac.uk/bloodbrain/). Although this tool is limited by its reliance on blood and brain samples from individuals 71 years of age and older, it does suggest that the differential methylation detected at chr5:172903876 in whole blood among the children whose mothers reported an infection prior to their pregnancy may reflect differential methylation in the child's brain as well. This would correspond to the second hypothesized model in our introduction for the relationship between blood and brain DNA methylation.

There are several challenges that these data present, including potential exposure misclassification due to the retrospective maternal self-report of infection. In SEED, exposure data were collected retrospectively at the time of enrollment, which means that mothers may have been recalling their pregnancy exposures up to five years after delivery. This raises the possibility of recall bias, which may or may not be differential by ASD case status, though we would not suspect differential recall based on methylation levels. Indeed, we saw strong associations between several categories of infection exposure and ASD risk in the main effect analysis; we are not able to rule out that these associations are due to recall bias, though mothers in SEED were asked about a number of different exposures during their pregnancies, and not just about infection.

We also know that self-reported infection exposure is not ideal and may not reflect the same underlying construct as an infection diagnosis recorded in the medical record. Prospectively collected information on infections meeting standard case definitions, along with serologic confirmation of infection, during periconception, pregnancy, and the early postpartum period would in this case be the gold standard. We do have access to prenatal and labor and delivery records for SEED mothers, but an analysis based on their medical records has yet to be performed; although this method of exposure assessment may also be limited, as mothers may have had infections or fevers that were treated at home and never noted in their medical record. There may be more significant methylation differences related to *in utero* infection exposure that we were not able to detect due to a non-ideal exposure assessment.

Table 4.18: Correlations between DNA methylation in whole blood and four brain regions for the three significant infection-exposure DMPs

Chr:position	exposure period	PFC-blood r (p-value) (n=74)	EC-blood r (p-value) (n=71)	STG-blood r (p-value) (n=75)	CER-blood r (p-value) (n=71)
chr5:172903876	T0	**0.873 (3.43e-24)**	**0.895 (7.62e-26)**	**0.894 (4.2e-27)**	**0.883 (2.57e-24)**
chr3:12947823	T3	0.181 (0.122)	0.268 (0.024)	0.00442 (0.97)	-0.197 (0.0999)
chr1:110306507	T3	-0.109 (0.353)	-0.302 (0.0104)	-0.0228 (0.846)	0.00835 (0.945)

PFC, prefrontal cortex; EC, entorhinal cortex; STG, superior temporal gyrus; CER, cerebellum

Bold denotes statistically significant correlation

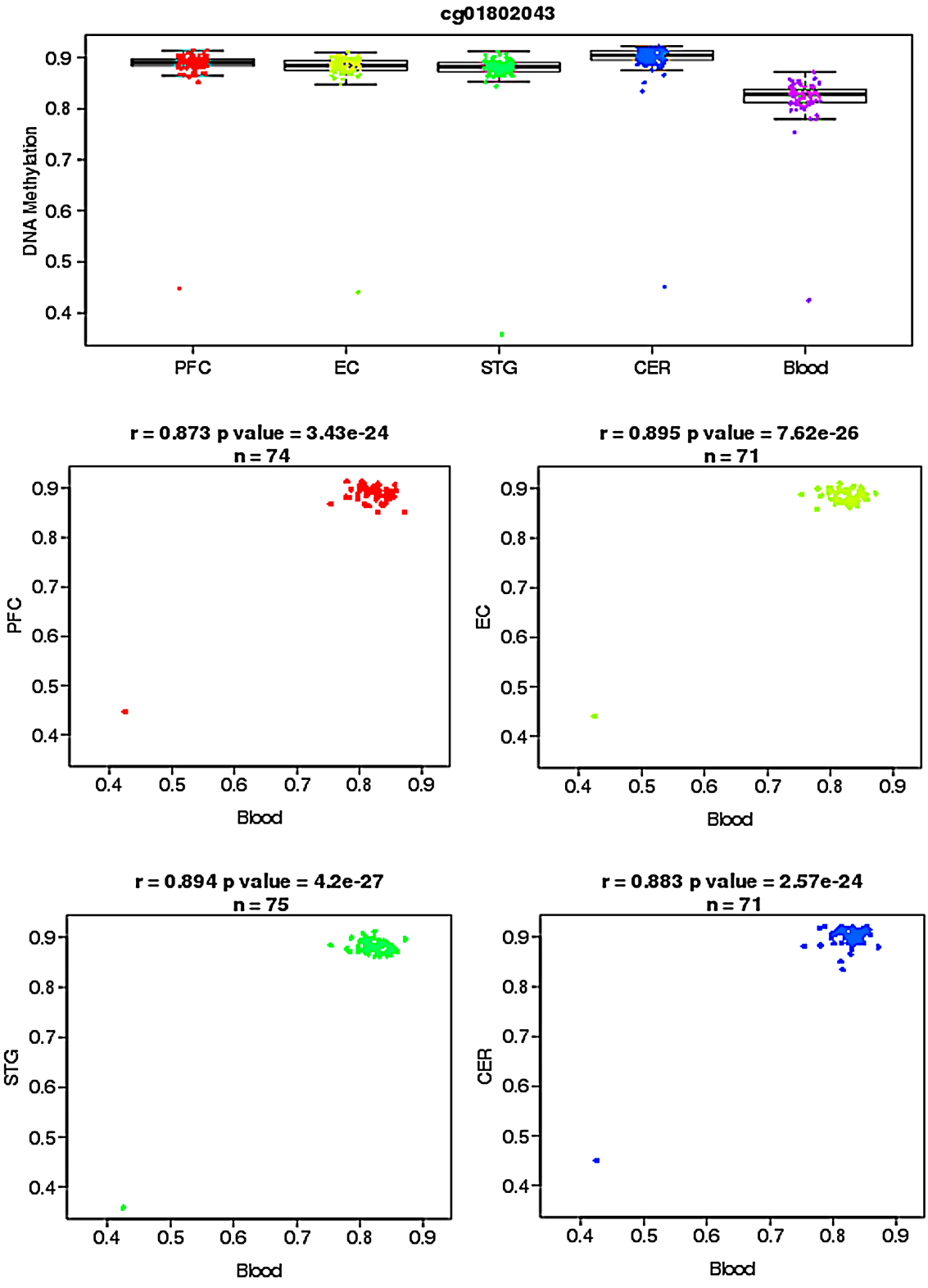

cg01802043
DNA Methylation
0.9
0.8
0.7
0.6
0.5
0.4
PFC
EC
STG
CER
Blood
r = 0.873 p value = 3.43e-24
n = 74
PFC
Blood
r = 0.895 p value = 7.62e-26
n = 71
EC
Blood
r = 0.894 p value = 4.2e-27
n = 75
STG
Blood
r = 0.883 p value = 2.57e-24
n = 71
CER
Blood

Figure 4.19: Correlation between methylation in whole blood at position chr5:172903876 and four brain regions, prefrontal cortex (PFC), entorhinal cortex (EC), superior temporal gyrus (STG) and cerebellum (CER), according to the Blood Brain DNA Methylation Comparison Tool (http://epigenetics.essex.ac.uk/bloodbrain/).

The derived variable available to us for this analysis collapsed many different infection types into variables reflecting exposure to any infection during specific time periods. The differential methylation that we found is thus detected based on a heterogeneous exposure; stronger associations between methylation and specific loci could exist when examining particular infection-methylation relationships. However, using more homogeneous exposure categories in the present study would have decreased our sample size and power to detect any association.

Multiple testing corrections is difficult to perform with epigenetic data, because of the known correlation between methylation at different probe sites. We did perform correction for multiple testing within each exposure category, calculating false discovery rates and performing a Bonferroni adjustment. However, we tested six different exposures, though these exposures are highly correlated (Table 4.19) and unlikely to be independently distributed. A strict Bonferroni cutoff (calculating the number of tests as six times the number of tested probes for each model) is likely to be inappropriate and overly conservative in this situation.

Table 4.19: Pearson's product-moment correlation between exposure variables

	T0	AP	T1	T2	T3	BF
Infection in 3 months prior to conception (T0)	1					
Infection at any time in pregnancy (AP)	0.20	1				
Trimester 1 (T1)	0.44	0.53	1			
Trimester 2 (T2)	0.28	0.63	0.37	1		
Trimester 3 (T3)	0.21	0.71	0.32	0.34	1	
Infection while breastfeeding (BF)	0.33	0.23	0.33	0.24	0.29	1

In our sample, we were unable to control for possible confounding or interaction with treatment for infection or fever. One previous study of the link between MIA exposure and ASD development did find an association between antibiotic use and ASD risk; however, they were unable to distinguish if this effect was observed because antibiotic use served as a proxy of MIA and infection, the true cause, or because it was independently contributing to disease (Atladottir, Henriksen et al. 2012). Another study found that anti-pyretics such as Tylenol actually attenuated an association between MIA and ASD (Zerbo, Qian et al. 2015). Unfortunately this kind of information is not

currently available to us. It does however, suggest future directions and extensions for our work.

Future work will include analyses to detect differential methylation in children who were exposed to maternal immune activation during gestation, using other markers or exposures of MIA. These will include fever during pregnancy, biomarkers of inflammation such a CRP or cytokines, and maternal autoimmune disease.

Chapter 5: Conclusions and future directions

This dissertation represents an effort to understand the biological consequences of exposure to an inflammatory environment *in utero* on a child's later development. First, we reviewed what is currently known about the genetic and environmental risk factors for Autism Spectrum Disorder (ASD), including prenatal exposure to maternal immune activation (**Chapter 1**). Then, we tested for an association between maternal immune activation during pregnancy and later development of ASD in the child (**Chapter 2**). We found that self-reported flu or genitourinary infections during pregnancy were not associated with an increased risk of the exposed child developing ASD, but that fever exposure was significantly associated. We found suggestive evidence that fever exposure during the third trimester might be particularly relevant for ASD risk, though we were limited by imprecise effect estimates due to sample size. The significant association between *in utero* exposure to fever and ASD risk was robust to sensitivity analyses for outcome and exposure misclassification.

Currently, epidemiologic studies of MIA have focused more on exposure to infections like influenza than fever itself. Our study is only the third we are aware of that looks specifically at ASD risk after exposure to maternal fever during gestation, and the first that looks in a predominantly low-income, black American population. We suggest that future studies of prenatal risk factors for ASD collect information on fever during

pregnancy, including more detailed information on fever timing, severity, duration, and treatment, including the use of anti-pyretics or antibiotics.

We also explored ways to maximize the information obtained from electronic medical records for research phenotypes, to benefit future research on ASD risk factors, including MIA exposure, with large clinical cohorts (**Chapter 3**). We were able to identify ICD-9-CM diagnoses that are most predictive of a child's score on an ASD screening questionnaire, the Social Communication Questionnaire (SCQ), using a machine learning technique called Random Forests (RF). As would be expected, the presence of the diagnosis code for current or active ASD (299.00) is the strongest predictor of SCQ score, but several other diagnoses provided useful information, predominantly conditions related to developmental delay, language disorder, and behavioral disorders. Incorporating information on these related diagnoses in ASD case definition within an EMR data set could improve the sensitivity and specificity of outcome classification. We thus explored the use of Latent Class Analysis (LCA), using as observed characteristics what RF identified as the most predictive ICD-9-CM diagnoses for SCQ score or the presence of a prior 299.00 diagnosis. LCA shows promise as a method that identifies false positive or negative ASD cases, though we anticipate future ASD evaluation data from the Boston Birth Cohort to formally assess sensitivity and specificity compared to a gold standard. Future directions for this work include validating the patterns observed with RF and LCA; the Boston Birth Cohort continues to increase the numbers of completed SCQ and SRS questionnaires among study participants. Additional SCQ results will allow us to test the model in a training data set,

while SRS results will allow us to test for ICD-9-CM diagnoses' predictive abilities for a quantitative autism phenotype.

Next, we explored the hypothesis that epigenetics plays an important role in the interface between environment and expression of the genome or disease risk, and tested whether there are detectable epigenetic alterations in the white blood cells of young children who were exposed to maternal immune activation *in utero (***Chapter 4**). We found evidence suggesting that epigenetic biomarkers of infection exposure—that occurred years prior, during gestation—can be detected in the whole blood of 2-5 year old children. Epigenetic changes resulting from prenatal exposures may be robust to postnatal exposures and persist through the first years of life, and perhaps even through adolescence. The set of loci or regions with differential methylation after exposure to prenatal infections could serve as a signature of past exposure regardless of disease status. This would be useful to exploit in future studies where we could predict exposure status based on measured DNA methylation levels, and then use it to test for primary associations between infection exposure and the outcome of interest.

We also found evidence that in this retrospective case-control study, maternal report of an infection during her pregnancy is associated with increased risk of ASD in her child. While this might be related to recall bias or differences in exposure assessment, as it is not consistent with the results from **Chapter 2** which had prospective data on infection exposure and asked separately about exposure to flu or genitourinary infections, it does raise the possibility that DNA methylation could be a mediator of a causal relationship between MIA and ASD.

References

Chapter 1

Anney, R., L. Klei, D. Pinto, J. Almeida, E. Bacchelli, G. Baird, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, J. Casey, J. Conroy, C. Correia, C. Corsello, E. L. Crawford, M. de Jonge, R. Delorme, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, J. Gilbert, C. Gillberg, J. T. Glessner, A. Green, J. Green, S. J. Guter, E. A. Heron, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, S. Jacob, G. P. Kenny, C. Kim, A. Kolevzon, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Law-Smith, M. Leboyer, A. Le Couteur, B. L. Leventhal, X. Q. Liu, F. Lombard, C. Lord, L. Lotspeich, S. C. Lund, T. R. Magalhaes, C. Mantoulan, C. J. McDougle, N. M. Melhem, A. Merikangas, N. J. Minshew, G. K. Mirza, J. Munson, C. Noakes, G. Nygren, K. Papanikolaou, A. T. Pagnamenta, B. Parrini, T. Paton, A. Pickles, D. J. Posey, F. Poustka, J. Ragoussis, R. Regan, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, S. Schlitt, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, V. Stoppioni, N. Sykes, R. Tancredi, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, J. A. Vorstman, S. Wallace, K. Wing, K. Wittemeyer, S. Wood, D. Zurawiecki, L. Zwaigenbaum, A. J. Bailey, A. Battaglia, R. M. Cantor, H. Coon, M. L. Cuccaro, G. Dawson, S. Ennis, C. M. Freitag, D. H. Geschwind, J. L. Haines, S. M. Klauck, W. M. McMahon, E. Maestrini, J. Miller, A. P. Monaco, S. F. Nelson, J. I. Nurnberger, Jr., G. Oliveira, J. R. Parr, M. A. Pericak-Vance, J. Piven, G. D. Schellenberg, S. W. Scherer, A. M. Vicente, T. H. Wassink, E. M. Wijsman, C. Betancur, J. D. Buxbaum, E. H. Cook, L. Gallagher, M. Gill, J. Hallmayer, A. D. Paterson, J. S. Sutcliffe, P. Szatmari, V. J. Vieland, H. Hakonarson and B. Devlin (2012). Individual common variants exert weak effects on the risk for autism spectrum disorders. Hum Mol Genet 21(21): 4781-4792.

Avella-Garcia, C. B., J. Julvez, J. Fortuny, C. Rebordosa, R. Garcia-Esteban, I. R. Galan, A. Tardon, C. L. Rodriguez-Bernal, C. Iniguez, A. Andiarena, L. Santa-Marina and J. Sunyer (2016). Acetaminophen use in pregnancy and neurodevelopment: attention function and autism spectrum symptoms. Int J Epidemiol. ePub.

Bailey, A., A. Le Couteur, I. Gottesman, P. Bolton, E. Simonoff, E. Yuzda and M. Rutter (1995). Autism as a strongly genetic disorder: evidence from a British twin study. Psychol Med 25(1): 63-77.

Bakulski, K. M., A. B. Singer and M. D. Fallin (2014). Genes and Environment in Autism Spectrum Disorders: An Integrated Perspective. Frontiers in Autism Research: pp. 335-374.

Barbaro, J. and C. Dissanayake (2010). Prospective identification of autism spectrum disorders in infancy and toddlerhood using developmental surveillance: the social attention and communication study. J Dev Behav Pediatr 31(5): 376-385.

Chauhan, N., M. S. Sen, S. Jhanda and S. Grover (2016). Psychiatric manifestations of congenital rubella syndrome: A case report and review of literature. J Pediatr Neurosci 11(2): 137-139.

Choi, G. B., Y. S. Yim, H. Wong, S. Kim, H. Kim, S. V. Kim, C. A. Hoeffer, D. R. Littman and J. R. Huh (2016). The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring. Science 351(6276): 933-939.

Christensen, D. L., J. Baio, K. Van Naarden Braun, D. Bilder, J. Charles, J. N. Constantino, J. Daniels, M. S. Durkin, R. T. Fitzgerald, M. Kurzius-Spencer, L. C. Lee, S. Pettygrove, C. Robinson, E. Schulz, C. Wells, M. S. Wingate, W. Zahorodny, M. Yeargin-Allsopp, C. Centers for Disease and Prevention (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years--Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. MMWR Surveill Summ 65(3): 1-23.

Conde-Agudelo, A., A. Rosas-Bermudez and M. H. Norton (2016). Birth Spacing and Risk of Autism and Other Neurodevelopmental Disabilities: A Systematic Review. Pediatrics 137(5).

Constantino, J. N., Y. Zhang, T. Frazier, A. M. Abbacchi and P. Law (2010). Sibling recurrence and the genetic epidemiology of autism. Am J Psychiatry 167(11): 1349-1356.

Courchesne, E., R. Carper and N. Akshoomoff (2003). Evidence of brain overgrowth in the first year of life in autism. JAMA 290(3): 337-344.

Croen, L. A., J. K. Grether, C. K. Yoshida, R. Odouli and V. Hendrick (2011). Antidepressant use during pregnancy and childhood autism spectrum disorders. Arch Gen Psychiatry 68(11): 1104-1112.

Dickerson, A. S., M. H. Rahbar, D. A. Pearson, R. S. Kirby, A. V. Bakian, D. A. Bilder, R. A. Harrington, S. Pettygrove, W. M. Zahorodny, L. A. Moye, 3rd, M. Durkin and M. Slay Wingate (2016). Autism spectrum disorder reporting in lower socioeconomic neighborhoods. Autism. ePub.

Fang, S. Y., S. Wang, N. Huang, H. H. Yeh and C. Y. Chen (2015). Prenatal Infection and Autism Spectrum Disorders in Childhood: A Population-Based Case-Control Study in Taiwan. Paediatr Perinat Epidemiol 29(4): 307-316.

Flanagan, J. E., R. Landa, A. Bhat and M. Bauman (2012). Head lag in infants at risk for autism: a preliminary study. Am J Occup Ther 66(5): 577-585.

Flores-Pajot, M. C., M. Ofner, M. T. Do, E. Lavigne and P. J. Villeneuve (2016). Childhood autism spectrum disorders and exposure to nitrogen dioxide, and particulate matter air pollution: A review and meta-analysis. Environ Res 151(763-776).

Folstein, S. and M. Rutter (1977). Infantile autism: a genetic study of 21 twin pairs. J Child Psychol Psychiatry 18(4): 297-321.

Froehlich-Santino, W., A. Londono Tobon, S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S. Ozonoff, C. Lajonchere, J. K. Grether, R. O'Hara and J. Hallmayer (2014). Prenatal and perinatal risk factors in a twin study of autism spectrum disorders. J Psychiatr Res 54(100-108).

Gao, Y., C. Sheng, R. H. Xie, W. Sun, E. Asztalos, D. Moddemann, L. Zwaigenbaum, M. Walker and S. W. Wen (2016). New Perspective on Impact of Folic Acid Supplementation during Pregnancy on Neurodevelopment/Autism in the Offspring Children - A Systematic Review. PLoS One 11(11): e0165626.

Garofoli, F., G. Lombardi, S. Orcesi, C. Pisoni, I. Mazzucchelli, M. Angelini, U. Balottin and M. Stronati (2017). An Italian Prospective Experience on the Association Between Congenital Cytomegalovirus Infection and Autistic Spectrum Disorder. J Autism Dev Disord. ePub.

Gaugler, T., L. Klei, S. J. Sanders, C. A. Bodea, A. P. Goldberg, A. B. Lee, M. Mahajan, D. Manaa, Y. Pawitan, J. Reichert, S. Ripke, S. Sandin, P. Sklar, O. Svantesson, A. Reichenberg, C. M. Hultman, B. Devlin, K. Roeder and J. D. Buxbaum (2014). Most genetic risk for autism resides with common variation. Nat Genet 46(8): 881-885.

Getahun, D., M. J. Fassett, M. R. Peltier, D. A. Wing, A. H. Xiang, V. Chiu and S. J. Jacobsen (2017). Association of Perinatal Risk Factors with Autism Spectrum Disorder. Am J Perinatol 34(3): 295-304.

Glatt, S. J., M. T. Tsuang, M. Winn, S. D. Chandler, M. Collins, L. Lopez, M. Weinfeld, C. Carter, N. Schork, K. Pierce and E. Courchesne (2012). Blood-based gene expression signatures of infants and toddlers with autism. J Am Acad Child Adolesc Psychiatry 51(9): 934-944 e932.

Glessner, J. T., K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, M. Imielinski, E. C. Frackelton, J. Reichert, E. L. Crawford, J. Munson, P. M. Sleiman, R. Chiavacci, K. Annaiah, K. Thomas, C. Hou, W. Glaberson, J. Flory, F. Otieno, M. Garris, L. Soorya, L. Klei, J. Piven, K. J. Meyer, E. Anagnostou, T. Sakurai, R. M. Game, D. S. Rudd, D. Zurawiecki, C. J. McDougle, L. K. Davis, J. Miller, D. J. Posey, S. Michaels, A. Kolevzon, J. M. Silverman, R. Bernier, S. E. Levy, R. T. Schultz, G. Dawson, T. Owley, W. M. McMahon, T. H. Wassink, J. A. Sweeney, J. I. Nurnberger, H. Coon, J. S. Sutcliffe, N. J. Minshew, S. F. Grant, M. Bucan, E. H. Cook, J.

D. Buxbaum, B. Devlin, G. D. Schellenberg and H. Hakonarson (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459(7246): 569-573.

Guinchat, V., P. Thorsen, C. Laurent, C. Cans, N. Bodeau and D. Cohen (2012). Pre-, peri- and neonatal risk factors for autism. Acta Obstet Gynecol Scand 91(3): 287-300.

Hadjkacem, I., H. Ayadi, M. Turki, S. Yaich, K. Khemekhem, A. Walha, L. Cherif, Y. Moalla and F. Ghribi (2016). Prenatal, perinatal and postnatal factors associated with autism spectrum disorder. J Pediatr (Rio J) 92(6): 595-601.

Hallmayer, J., S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S. Ozonoff, C. Lajonchere, J. K. Grether and N. Risch (2011). Genetic heritability and shared environmental factors among twin pairs with autism. Arch Gen Psychiatry 68(11): 1095-1102.

Harrington, R. A., L. C. Lee, R. M. Crum, A. W. Zimmerman and I. Hertz-Picciotto (2014). Prenatal SSRI use and offspring with autism spectrum disorder or developmental delay. Pediatrics 133(5): e1241-1248.

Hazlett, H. C., H. Gu, B. C. Munsell, S. H. Kim, M. Styner, J. J. Wolff, J. T. Elison, M. R. Swanson, H. Zhu, K. N. Botteron, D. L. Collins, J. N. Constantino, S. R. Dager, A. M. Estes, A. C. Evans, V. S. Fonov, G. Gerig, P. Kostopoulos, R. C. McKinstry, J. Pandey, S. Paterson, J. R. Pruett, R. T. Schultz, D. W. Shaw, L. Zwaigenbaum, J. Piven, I. Network, S. Clinical, C. Data Coordinating, C. Image Processing and A. Statistical (2017). Early brain development in infants at high risk for autism spectrum disorder. Nature 542(7641): 348-351.

Iossifov, I., M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Rosenbaum, B. Yamrom, Y. H. Lee, G. Narzisi, A. Leotta, J. Kendall, E. Grabowska, B. Ma, S. Marks, L. Rodgers, A. Stepansky, J. Troge, P. Andrews, M. Bekritsky, K. Pradhan, E. Ghiban, M. Kramer, J. Parla, R. Demeter, L. L. Fulton, R. S. Fulton, V. J. Magrini, K. Ye, J. C. Darnell, R. B. Darnell, E. R. Mardis, R. K. Wilson, M. C. Schatz, W. R. McCombie and M. Wigler (2012). De novo gene disruptions in children on the autistic spectrum. Neuron 74(2): 285-299.

Jiang, H. Y., L. L. Xu, L. Shao, R. M. Xia, Z. H. Yu, Z. X. Ling, F. Yang, M. Deng and B. Ruan (2016). Maternal infection during pregnancy and risk of autism spectrum disorders: A systematic review and meta-analysis. Brain Behav Immun 58(165-172).

Jones, K. L., L. A. Croen, C. K. Yoshida, L. Heuer, R. Hansen, O. Zerbo, G. N. DeLorenze, M. Kharrazi, R. Yolken, P. Ashwood and J. Van de Water (2017). Autism with intellectual disability is associated with increased levels of maternal cytokines and chemokines during gestation. Mol Psychiatry 22(2): 273-279.

Joseph, R. M., S. J. Korzeniewski, E. N. Allred, T. M. O'Shea, T. Heeren, J. A. Frazier, J. Ware, D. Hirtz, A. Leviton, K. Kuban and E. S. Investigators (2017). Extremely low gestational age and very low birthweight for gestational age are risk factors for autism spectrum disorder in a large cohort study of 10-year-old children born at 23-27 weeks' gestation. Am J Obstet Gynecol 216(3): 304 e301-304 e316.

Klei, L., S. J. Sanders, M. T. Murtha, V. Hus, J. K. Lowe, A. J. Willsey, D. Moreno-De-Luca, T. W. Yu, E. Fombonne, D. Geschwind, D. E. Grice, D. H. Ledbetter, C. Lord, S. M. Mane, C. L. Martin, D. M. Martin, E. M. Morrow, C. A. Walsh, N. M. Melhem, P. Chaste, J. S. Sutcliffe, M. W. State, E. H. Cook, Jr., K. Roeder and B. Devlin (2012). Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism 3(1): 9.

Labouesse, M. A., E. Dong, D. R. Grayson, A. Guidotti and U. Meyer (2015). Maternal immune activation induces GAD1 and GAD2 promoter remodeling in the offspring prefrontal cortex. Epigenetics 10(12): 1143-1155.

Ladd-Acosta, C., C. Shu, B. K. Lee, N. Gidaya, A. Singer, L. A. Schieve, D. E. Schendel, N. Jones, J. L. Daniels, G. C. Windham, C. J. Newschaffer, L. A. Croen, A. P. Feinberg and M. Daniele Fallin (2016). Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environ Res 144(Pt A): 139-148.

Lam, J., P. Sutton, A. Kalkbrenner, G. Windham, A. Halladay, E. Koustas, C. Lawler, L. Davidson, N. Daniels, C. Newschaffer and T. Woodruff (2016). A Systematic Review and Meta-Analysis of Multiple Airborne Pollutants and Autism Spectrum Disorder. PLoS One 11(9): e0161851.

Li, M., M. D. Fallin, A. Riley, R. Landa, S. O. Walker, M. Silverstein, D. Caruso, C. Pearson, S. Kiang, J. L. Dahm, X. Hong, G. Wang, M. C. Wang, B. Zuckerman and X. Wang (2016). The Association of Maternal Obesity and Diabetes With Autism and Other Developmental Disabilities. Pediatrics 137(2): e20152206.

Liew, Z., B. Ritz, J. Virk and J. Olsen (2016). Maternal use of acetaminophen during pregnancy and risk of autism spectrum disorders in childhood: A Danish national birth cohort study. Autism Res 9(9): 951-958.

Logan, J. W., O. Dammann, E. N. Allred, C. Dammann, K. Beam, R. M. Joseph, T. M. O'Shea, A. Leviton, K. C. Kuban and E. S. Investigators (2017). Early postnatal illness severity scores predict neurodevelopmental impairments at 10 years of age in children born extremely preterm. J Perinatol. ePub.

Loke, Y. J., A. J. Hannan and J. M. Craig (2015). The Role of Epigenetic Change in Autism Spectrum Disorders. Front Neurol 6(107).

Lyall, K., L. A. Croen, A. Sjodin, C. K. Yoshida, O. Zerbo, M. Kharrazi and G. C. Windham (2017). Polychlorinated Biphenyl and Organochlorine Pesticide Concentrations

in Maternal Mid-Pregnancy Serum Samples: Association with Autism Spectrum Disorder and Intellectual Disability. Environ Health Perspect 125(3): 474-480.

Magnusson, C., M. Lundberg, B. K. Lee, D. Rai, H. Karlsson, R. Gardner, K. Kosidou, S. Arver and C. Dalman (2016). Maternal vitamin D deficiency and the risk of autism spectrum disorders: population-based study. BJPsych Open 2(2): 170-172.

Mahic, M., S. Mjaaland, H. M. Bovelstad, N. Gunnes, E. Susser, M. Bresnahan, A. S. Oyen, B. Levin, X. Che, D. Hirtz, T. Reichborn-Kjennerud, S. Schjolberg, C. Roth, P. Magnus, C. Stoltenberg, P. Suren, M. Hornig and W. I. Lipkin (2017). Maternal Immunoreactivity to Herpes Simplex Virus 2 and Risk of Autism Spectrum Disorder in Male Offspring. mSphere 2(1): e00016-17.

Mazumdar, S., A. Winter, K. Y. Liu and P. Bearman (2013). Spatial clusters of autism births and diagnoses point to contextual drivers of increased prevalence. Soc Sci Med 95: 87-96.

Meyer, U., J. Feldon and O. Dammann (2011). Schizophrenia and autism: both shared and disorder-specific pathogenesis via perinatal inflammation? Pediatr Res 69(5 Pt 2): 26R-33R.

Neale, B. M., Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C. F. Lin, C. Stevens, L. S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, Y. Wu, L. Lewis, Y. Han, B. F. Voight, E. Lim, E. Rossin, A. Kirby, J. Flannick, M. Fromer, K. Shakir, T. Fennell, K. Garimella, E. Banks, R. Poplin, S. Gabriel, M. DePristo, J. R. Wimbish, B. E. Boone, S. E. Levy, C. Betancur, S. Sunyaev, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, Jr., B. Devlin, R. A. Gibbs, K. Roeder, G. D. Schellenberg, J. S. Sutcliffe and M. J. Daly (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485(7397): 242-245.

O'Roak, B. J., L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure and E. E. Eichler (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485(7397): 246-250.

Ozonoff, S., G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Dobkins, T. Hutman, J. M. Iverson, R. Landa, S. J. Rogers, M. Sigman and W. L. Stone (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. Pediatrics 128(3): e488-495.

Pang, Y., X. Dai, A. Roller, K. Carter, I. Paul, A. J. Bhatt, R. C. Lin and L. W. Fan (2016). Early Postnatal Lipopolysaccharide Exposure Leads to Enhanced Neurogenesis and Impaired Communicative Functions in Rats. PLoS One 11(10): e0164403.

Papadakis, A. I., D. Baltzis, R. C. Buensuceso, P. Peidis and A. E. Koromilas (2011). Development of transgenic mice expressing a conditionally active form of the eIF2alpha kinase PKR. Genesis 49(9): 743-749.

Pediatrics, A. A. o. (2017). "“Vaccine Safety: Examine the Evidence”." Retrieved March 14, 2017, from https://www.healthychildren.org/English/safety-prevention/immunizations/Pages/Vaccine-Studies-Examine-the-Evidence.aspx.

Peterson, K. and P. Barbel (2013). On alert for autism spectrum disorders. Nursing 43(4): 28-34; quiz 35.

Pramparo, T., K. Pierce, M. V. Lombardo, C. Carter Barnes, S. Marinero, C. Ahrens-Barbeau, S. S. Murray, L. Lopez, R. Xu and E. Courchesne (2015). Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. JAMA Psychiatry 72(4): 386-394.

Rai, D., B. K. Lee, C. Dalman, J. Golding, G. Lewis and C. Magnusson (2013). Parental depression, maternal antidepressant use during pregnancy, and risk of autism spectrum disorders: population based case-control study. BMJ 346(f2059).

Richetto, J., R. Massart, U. Weber-Stadlbauer, M. Szyf, M. A. Riva and U. Meyer (2017). Genome-wide DNA Methylation Changes in a Mouse Model of Infection-Mediated Neurodevelopmental Disorders. Biol Psychiatry 81(3): 265-276.

Roberts, E. M., P. B. English, J. K. Grether, G. C. Windham, L. Somberg and C. Wolff (2007). Maternal residence near agricultural pesticide applications and autism spectrum disorders among children in the California Central Valley. Environ Health Perspect 115(10): 1482-1489.

Robinson, E. B., B. St Pourcain, V. Anttila, J. A. Kosmicki, B. Bulik-Sullivan, J. Grove, J. Maller, K. E. Samocha, S. J. Sanders, S. Ripke, J. Martin, M. V. Hollegaard, T. Werge, D. M. Hougaard, P.-S. S. I. B. A. G. i, B. M. Neale, D. M. Evans, D. Skuse, P. B. Mortensen, A. D. Borglum, A. Ronald, G. D. Smith and M. J. Daly (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nat Genet 48(5): 552-555.

Sanders, S. J., A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, C. E. Mason, K. Bilguvar, P. B. Celestino-Soper, M. Choi, E. L. Crawford, L. Davis, N. R. Wright, R. M. Dhodapkar, M. DiCola, N. M. DiLullo, T. V. Fernandez, V. Fielding-Singh, D. O. Fishman, S. Frahm, R. Garagaloyan, G. S. Goh, S. Kammela, L. Klei, J. K. Lowe, S. C. Lund, A. D. McGrew, K. A. Meyer, W. J. Moffat, J. D. Murdoch, B. J. O'Roak, G. T. Ober, R. S. Pottenger, M. J.

Raubeson, Y. Song, Q. Wang, B. L. Yaspan, T. W. Yu, I. R. Yurkiewicz, A. L. Beaudet, R. M. Cantor, M. Curland, D. E. Grice, M. Gunel, R. P. Lifton, S. M. Mane, D. M. Martin, C. A. Shaw, M. Sheldon, J. A. Tischfield, C. A. Walsh, E. M. Morrow, D. H. Ledbetter, E. Fombonne, C. Lord, C. L. Martin, A. I. Brooks, J. S. Sutcliffe, E. H. Cook, Jr., D. Geschwind, K. Roeder, B. Devlin and M. W. State (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70(5): 863-885.

Sanders, S. J., M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. A. Meyer, K. Bilguvar, S. M. Mane, N. Sestan, R. P. Lifton, M. Gunel, K. Roeder, D. H. Geschwind, B. Devlin and M. W. State (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485(7397): 237-241.

Sandin, S., P. Lichtenstein, R. Kuja-Halkola, H. Larsson, C. M. Hultman and A. Reichenberg (2014). The familial risk of autism. JAMA 311(17): 1770-1777.

Sandin, S., D. Schendel, P. Magnusson, C. Hultman, P. Suren, E. Susser, T. Gronborg, M. Gissler, N. Gunnes, R. Gross, M. Henning, M. Bresnahan, A. Sourander, M. Hornig, K. Carter, R. Francis, E. Parner, H. Leonard, M. Rosanoff, C. Stoltenberg and A. Reichenberg (2016). Autism risk associated with parental age and with increasing difference in age between the parents. Mol Psychiatry 21(5): 693-700.

Schmidt, R. J., K. Lyall and I. Hertz-Picciotto (2014). Environment and Autism: Current State of the Science. Cut Edge Psychiatry Pract 1(4): 21-38.

Schmidt, R. J., D. J. Tancredi, S. Ozonoff, R. L. Hansen, J. Hartiala, H. Allayee, L. C. Schmidt, F. Tassone and I. Hertz-Picciotto (2012). Maternal periconceptional folic acid intake and risk of autism spectrum disorders and developmental delay in the CHARGE (CHildhood Autism Risks from Genetics and Environment) case-control study. Am J Clin Nutr 96(1): 80-89.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye and M. Wigler (2007). Strong association of de novo copy number mutations with autism. Science 316(5823): 445-449.

Shelton, J. F., E. M. Geraghty, D. J. Tancredi, L. D. Delwiche, R. J. Schmidt, B. Ritz, R. L. Hansen and I. Hertz-Picciotto (2014). Neurodevelopmental disorders and prenatal residential proximity to agricultural pesticides: the CHARGE study. Environ Health Perspect 122(10): 1103-1109.

Smith, N. W., G. M. Strutton, M. D. Walsh, G. R. Wright, G. J. Seymour, M. F. Lavin and R. A. Gardiner (1990). Transferrin receptor expression in primary superficial human bladder tumours identifies patients who develop recurrences. Br J Urol 65(4): 339-344.

Steffenburg, S., C. Gillberg, L. Hellgren, L. Andersson, I. C. Gillberg, G. Jakobsson and M. Bohman (1989). A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. J Child Psychol Psychiatry 30(3): 405-416.

Taniai, H., T. Nishiyama, T. Miyachi, M. Imaeda and S. Sumi (2008). Genetic influences on the broad spectrum of autism: study of proband-ascertained twins. Am J Med Genet B Neuropsychiatr Genet 147B(6): 844-849.

Tick, B., P. Bolton, F. Happe, M. Rutter and F. Rijsdijk (2016). Heritability of autism spectrum disorders: a meta-analysis of twin studies. J Child Psychol Psychiatry 57(5): 585-595.

Volk, H. E., F. Lurmann, B. Penfold, I. Hertz-Picciotto and R. McConnell (2013). Traffic-related air pollution, particulate matter, and autism. JAMA Psychiatry 70(1): 71-77.

Wang, K., H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong, T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg and H. Hakonarson (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459(7246): 528-533.

Weiss, L. A., Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B. L. Wu, M. J. Daly and C. Autism (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358(7): 667-675.

Werling, D. M. and D. H. Geschwind (2013). Sex differences in autism spectrum disorders. Curr Opin Neurol 26(2): 146-153.

Zerbo, O., A. M. Iosif, C. Walker, S. Ozonoff, R. L. Hansen and I. Hertz-Picciotto (2013). Is maternal influenza or fever during pregnancy associated with autism or developmental delays? Results from the CHARGE (CHildhood Autism Risks from Genetics and Environment) study. J Autism Dev Disord 43(1): 25-33.

Zerbo, O., C. Yoshida, E. P. Gunderson, K. Dorward and L. A. Croen (2015). Interpregnancy Interval and Risk of Autism Spectrum Disorders. Pediatrics 136(4): 651-657.

Chapter 2

Atladottir, H. O., T. B. Henriksen, D. E. Schendel and E. T. Parner (2012). Autism after infection, febrile episodes, and antibiotic use during pregnancy: an exploratory study. Pediatrics 130(6): e1447-1454.

Atladottir, H. O., P. Thorsen, L. Ostergaard, D. E. Schendel, S. Lemcke, M. Abdallah and E. T. Parner (2010). Maternal infection requiring hospitalization during pregnancy and autism spectrum disorders. J Autism Dev Disord 40(12): 1423-1430.

Barouki, R., P. D. Gluckman, P. Grandjean, M. Hanson and J. J. Heindel (2012). Developmental origins of non-communicable disease: implications for research and public health. Environ Health 11: 42.

Bauman, M. D., A. M. Iosif, S. E. Smith, C. Bregere, D. G. Amaral and P. H. Patterson (2014). Activation of the maternal immune system during pregnancy alters behavioral development of rhesus monkey offspring. Biol Psychiatry 75(4): 332-341.

Christensen, D. L., J. Baio, K. Van Naarden Braun, D. Bilder, J. Charles, J. N. Constantino, J. Daniels, M. S. Durkin, R. T. Fitzgerald, M. Kurzius-Spencer, L. C. Lee, S. Pettygrove, C. Robinson, E. Schulz, C. Wells, M. S. Wingate, W. Zahorodny, M. Yeargin-Allsopp, C. Centers for Disease and Prevention (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years--Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. MMWR Surveill Summ 65(3): 1-23.

Collier, S. A., S. A. Rasmussen, M. L. Feldkamp, M. A. Honein and S. National Birth Defects Prevention (2009). Prevalence of self-reported infection during pregnancy among control mothers in the National Birth Defects Prevention Study. Birth Defects Res A Clin Mol Teratol 85(3): 193-201.

Dodds, L., A. Spencer, S. Shea, D. Fell, B. A. Armson, A. C. Allen and S. Bryson (2009). Validity of autism diagnoses using administrative health data. Chronic Dis Can 29(3): 102-107.

Dreier, J. W., A. M. Andersen and G. Berg-Beckhoff (2014). Systematic review and meta-analyses: fever in pregnancy and health impacts in the offspring. Pediatrics 133(3): e674-688.

Fang, S. Y., S. Wang, N. Huang, H. H. Yeh and C. Y. Chen (2015). Prenatal Infection and Autism Spectrum Disorders in Childhood: A Population-Based Case-Control Study in Taiwan. Paediatr Perinat Epidemiol 29(4): 307-316.

Hallmayer, J., S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S. Ozonoff, C. Lajonchere, J. K.

Grether and N. Risch (2011). Genetic heritability and shared environmental factors among twin pairs with autism. Arch Gen Psychiatry 68(11): 1095-1102.

Ho, D. E., K. Imai, G. King and S. E.A. (2007). Matching as nonparametric preprocessing for reducing model dependance in parametric causal inference. Political Analysis 15: 199-236.

Ho, D. E., K. Imai, G. King and S. E.A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. Journal of Statistical Software 42(8).

Hsiao, E. Y., S. W. McBride, J. Chow, S. K. Mazmanian and P. H. Patterson (2012). Modeling an autism risk factor in mice leads to permanent immune dysregulation. Proc Natl Acad Sci U S A 109(31): 12776-12781.

Jiang, H. Y., L. L. Xu, L. Shao, R. M. Xia, Z. H. Yu, Z. X. Ling, F. Yang, M. Deng and B. Ruan (2016). Maternal infection during pregnancy and risk of autism spectrum disorders: A systematic review and meta-analysis. Brain Behav Immun 58: 165-172.

Lash T.L, Fox M.P, and Fink A.K. "Applying Quantitative Bias Analysis to Epidemiologic Data". ('Springer', 2009).

Le Belle, J. E., J. Sperry, A. Ngo, Y. Ghochani, D. R. Laks, M. Lopez-Aranda, A. J. Silva and H. I. Kornblum (2014). Maternal inflammation contributes to brain overgrowth and autism-associated behaviors through altered redox signaling in stem and progenitor cells. Stem Cell Reports 3(5): 725-734.

Lee, B. K., C. Magnusson, R. M. Gardner, A. Blomstrom, C. J. Newschaffer, I. Burstyn, H. Karlsson and C. Dalman (2015). Maternal hospitalization with infection during pregnancy and risk of autism spectrum disorders. Brain Behav Immun 44: 100-105.

Liew, Z., B. Ritz, J. Virk and J. Olsen (2016). Maternal use of acetaminophen during pregnancy and risk of autism spectrum disorders in childhood: A Danish national birth cohort study. Autism Res 9(9): 951-958.

Lyall, K., R. J. Schmidt and I. Hertz-Picciotto (2014). Maternal lifestyle and environmental risk factors for autism spectrum disorders. Int J Epidemiol 43(2): 443-464.

Machado, C. J., A. M. Whitaker, S. E. Smith, P. H. Patterson and M. D. Bauman (2015). Maternal immune activation in nonhuman primates alters social attention in juvenile offspring. Biol Psychiatry 77(9): 823-832.

Malkova, N. V., C. Z. Yu, E. Y. Hsiao, M. J. Moore and P. H. Patterson (2012). Maternal immune activation yields offspring displaying mouse versions of the three core symptoms of autism. Brain Behav Immun 26(4): 607-616.

Marques, A. H., T. G. O'Connor, C. Roth, E. Susser and A. L. Bjorke-Monsen (2013). The influence of maternal prenatal and early childhood nutrition and maternal prenatal stress on offspring immune system development and neurodevelopmental disorders. Front Neurosci 7: 120.

Matcovitch-Natan, O., D. R. Winter, A. Giladi, S. Vargas Aguilar, A. Spinrad, S. Sarrazin, H. Ben-Yehuda, E. David, F. Zelada Gonzalez, P. Perrin, H. Keren-Shaul, M. Gury, D. Lara-Astaiso, C. A. Thaiss, M. Cohen, K. Bahar Halpern, K. Baruch, A. Deczkowska, E. Lorenzo-Vivas, S. Itzkovitz, E. Elinav, M. H. Sieweke, M. Schwartz and I. Amit (2016). Microglia development follows a stepwise program to regulate brain homeostasis. Science 353(6301): aad8670.

Miller, V. M., Y. Zhu, C. Bucher, W. McGinnis, L. K. Ryan, A. Siegel and S. Zalcman (2013). Gestational flu exposure induces changes in neurochemicals, affiliative hormones and brainstem inflammation, in addition to autism-like behaviors in mice. Brain Behav Immun 33: 153-163.

Ohkawara, T., T. Katsuyama, M. Ida-Eto, N. Narita and M. Narita (2015). Maternal viral infection during pregnancy impairs development of fetal serotonergic neurons. Brain Dev 37(1): 88-93.

Persico, A. M. and V. Napolioni (2013). Autism genetics. Behav Brain Res 251: 95-112.

Rice, D. and S. Barone, Jr. (2000). Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. Environ Health Perspect 108 Suppl 3: 511-533.

Rodier, P. M., J. L. Ingram, B. Tisdale, S. Nelson and J. Romano (1996). Embryological origin for autism: developmental anomalies of the cranial nerve motor nuclei. J Comp Neurol 370(2): 247-261.

Sandin, S., P. Lichtenstein, R. Kuja-Halkola, H. Larsson, C. M. Hultman and A. Reichenberg (2014). The familial risk of autism. JAMA 311(17): 1770-1777.

Schlotz, W. and D. I. Phillips (2009). Fetal origins of mental health: evidence and mechanisms. Brain Behav Immun 23(7): 905-916.

Schwartzer, J. J., M. Careaga, C. E. Onore, J. A. Rushakoff, R. F. Berman and P. Ashwood (2013). Maternal immune activation and strain specific interactions in the development of autism-like behaviors in mice. Transl Psychiatry 3: e240.

Shi, L., S. E. Smith, N. Malkova, D. Tse, Y. Su and P. H. Patterson (2009). Activation of the maternal immune system alters cerebellar development in the offspring. Brain Behav Immun 23(1): 116-123.

Stoner, R., M. L. Chow, M. P. Boyle, S. M. Sunkin, P. R. Mouton, S. Roy, A. Wynshaw-Boris, S. A. Colamarino, E. S. Lein and E. Courchesne (2014). Patches of disorganization in the neocortex of children with autism. N Engl J Med 370(13): 1209-1219.

Straley, M. E., K. L. Togher, A. M. Nolan, L. C. Kenny and G. W. O'Keeffe (2014). LPS alters placental inflammatory and endocrine mediators and inhibits fetal neurite growth in affected offspring during late gestation. Placenta 35(8): 533-538.

Wang, G., S. Divall, S. Radovick, D. Paige, Y. Ning, Z. Chen, Y. Ji, X. Hong, S. O. Walker, D. Caruso, C. Pearson, M. C. Wang, B. Zuckerman, T. L. Cheng and X. Wang (2014). Preterm birth and random plasma insulin levels at birth and in early childhood. JAMA 311(6): 587-596.

Wang, X., B. Zuckerman, C. Pearson, G. Kaufman, C. Chen, G. Wang, T. Niu, P. H. Wise, H. Bauchner and X. Xu (2002). Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. JAMA 287(2): 195-202.

Zerbo, O., A. M. Iosif, C. Walker, S. Ozonoff, R. L. Hansen and I. Hertz-Picciotto (2013). Is maternal influenza or fever during pregnancy associated with autism or developmental delays? Results from the CHARGE (CHildhood Autism Risks from Genetics and Environment) study. J Autism Dev Disord 43(1): 25-33.

Zerbo, O., Y. Qian, C. Yoshida, B. H. Fireman, N. P. Klein and L. A. Croen (2017). Association Between Influenza Infection and Vaccination During Pregnancy and Risk of Autism Spectrum Disorder. JAMA Pediatr 171(1): e163609.

Zerbo, O., Y. Qian, C. Yoshida, J. K. Grether, J. Van de Water and L. A. Croen (2015). Maternal Infection During Pregnancy and Autism Spectrum Disorders. J Autism Dev Disord 45(12): 4015-4025.

Chapter 3

Aldinger, K. A., C. J. Lane, J. Veenstra-VanderWeele and P. Levitt (2015). Patterns of Risk for Multiple Co-Occurring Medical Conditions Replicate Across Distinct Cohorts of Children with Autism Spectrum Disorder. Autism Res 8(6): 771-781.

Bernatsky, S., L. Lix, J. G. Hanly, M. Hudson, E. Badley, C. Peschken, C. A. Pineau, A. E. Clarke, P. R. Fortin, M. Smith, P. Belisle, C. Lagace, L. Bergeron and L. Joseph (2011). Surveillance of systemic autoimmune rheumatic diseases using administrative data. Rheumatol Int 31(4): 549-554.

Bowman, S., R. M. Cleland and S. Staggs (2015). A Strategic Plan for Integrating ICD-10 in Your Practice and Workflow. Am Soc Clin Oncol Educ Booke91-98.

Breiman, L. Random Forests. Machine Learning 45(1): 5-32.

Breiman L, F. J., Olshen RA, Stone CJ (1984). Classification and Regression Trees, CRC Press.

Chandler, S., T. Charman, G. Baird, E. Simonoff, T. Loucas, D. Meldrum, M. Scott and A. Pickles (2007). Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders. J Am Acad Child Adolesc Psychiatry 46(10): 1324-1332.

Christensen, D. L., J. Baio, K. Van Naarden Braun, D. Bilder, J. Charles, J. N. Constantino, J. Daniels, M. S. Durkin, R. T. Fitzgerald, M. Kurzius-Spencer, L. C. Lee, S. Pettygrove, C. Robinson, E. Schulz, C. Wells, M. S. Wingate, W. Zahorodny, M. Yeargin-Allsopp, C. Centers for Disease and Prevention (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years--Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. MMWR Surveill Summ 65(3): 1-23.

Clifton, D. A., K. E. Niehaus, P. Charlton and G. W. Colopy (2015). Health Informatics via Machine Learning for the Clinical Management of Patients. Yearb Med Inform 10(1): 38-43.

Close, H. A., L. C. Lee, C. N. Kaufmann and A. W. Zimmerman (2012). Co-occurring conditions and change in diagnosis in autism spectrum disorders. Pediatrics 129(2): e305-316.

Coleman, K. J., M. A. Lutsky, V. Yau, Y. Qian, M. E. Pomichowski, P. M. Crawford, F. L. Lynch, J. M. Madden, A. Owen-Smith, J. A. Pearson, K. A. Pearson, D. Rusinak, V. P. Quinn and L. A. Croen (2015). Validation of Autism Spectrum Disorder Diagnoses in Large Healthcare Systems with Electronic Medical Records. J Autism Dev Disord 45(7): 1989-1996.

Collins, L. M. and S. T. Lanza (2010). Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences. Hoboken, NJ, John Wiley & Sons, Inc.

Connolly, N., J. Anixt, P. Manning, I. L. D. Ping, K. A. Marsolo and K. Bowers (2016). Maternal metabolic risk factors for autism spectrum disorder-An analysis of electronic medical records and linked birth data. Autism Res 9(8): 829-837.

Constantino, J. N. (2011). The quantitative nature of autistic social impairment. Pediatr Res 69(5 Pt 2): 55R-62R.

Constantino, J. N., S. A. Davis, R. D. Todd, M. K. Schindler, M. M. Gross, S. L. Brophy, L. M. Metzger, C. S. Shoushtari, R. Splinter and W. Reich (2003). Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. J Autism Dev Disord 33(4): 427-433.

Crawford, D. C., D. R. Crosslin, G. Tromp, I. J. Kullo, H. Kuivaniemi, M. G. Hayes, J. C. Denny, W. S. Bush, J. L. Haines, D. M. Roden, C. A. McCarty, G. P. Jarvik and M. D. Ritchie (2014). eMERGEing progress in genomics-the first seven years. Front Genet 5(184).

Denny, J. C., L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, M. A. Basford, D. S. Carrell, P. L. Peissig, A. N. Kho, J. A. Pacheco, L. V. Rasmussen, D. R. Crosslin, P. K. Crane, J. Pathak, S. J. Bielinski, S. A. Pendergrass, H. Xu, L. A. Hindorff, R. Li, T. A. Manolio, C. G. Chute, R. L. Chisholm, E. B. Larson, G. P. Jarvik, M. H. Brilliant, C. A. McCarty, I. J. Kullo, J. L. Haines, D. C. Crawford, D. R. Masys and D. M. Roden (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 31(12): 1102-1110.

Dodds, L., A. Spencer, S. Shea, D. Fell, B. A. Armson, A. C. Allen and S. Bryson (2009). Validity of autism diagnoses using administrative health data. Chronic Dis Can 29(3): 102-107.

Doshi-Velez, F., Y. Ge and I. Kohane (2014). Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics 133(1): e54-63.

Eaves, L. C., H. D. Wingert, H. H. Ho and E. C. Mickelson (2006). Screening for autism spectrum disorders with the social communication questionnaire. J Dev Behav Pediatr 27(2 Suppl): S95-S103.

Gillberg, C. (2010). The ESSENCE in child psychiatry: Early Symptomatic Syndromes Eliciting Neurodevelopmental Clinical Examinations. Res Dev Disabil 31(6): 1543-1551.

Khalilia, M., S. Chakraborty and M. Popescu (2011). Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making 11(1): 1-13.

Kohane, I. S., A. McMurry, G. Weber, D. MacFadden, L. Rappaport, L. Kunkel, J. Bickel, N. Wattanasin, S. Spence, S. Murphy and S. Churchill (2012). The co-morbidity burden of children and young adults with autism spectrum disorders. PLoS One 7(4): e33224.

Levy, S. E., E. Giarelli, L. C. Lee, L. A. Schieve, R. S. Kirby, C. Cunniff, J. Nicholas, J. Reaven and C. E. Rice (2010). Autism spectrum disorder and co-occurring developmental, psychiatric, and medical conditions among children in multiple populations of the United States. J Dev Behav Pediatr 31(4): 267-275.

Lingren, T., P. Chen, J. Bochenek, F. Doshi-Velez, P. Manning-Courtney, J. Bickel, L. Wildenger Welchons, J. Reinhold, N. Bing, Y. Ni, W. Barbaresi, F. Mentch, M. Basford, J. Denny, L. Vazquez, C. Perry, B. Namjou, H. Qiu, J. Connolly, D. Abrams, I. A. Holm, B. A. Cobb, N. Lingren, I. Solti, H. Hakonarson, I. S. Kohane, J. Harley and G. Savova (2016). Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. PLoS One 11(7): e0159621.

Maenner, M. J., M. Yeargin-Allsopp, K. Van Naarden Braun, D. L. Christensen and L. A. Schieve (2016). Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder. PLoS One 11(12): e0168224.

Mazumdar, S., A. Winter, K. Y. Liu and P. Bearman (2013). Spatial clusters of autism births and diagnoses point to contextual drivers of increased prevalence. Soc Sci Med 95(87-96).

Namjou, B., K. Marsolo, R. J. Caroll, J. C. Denny, M. D. Ritchie, S. S. Verma, T. Lingren, A. Porollo, B. L. Cobb, C. Perry, L. C. Kottyan, M. E. Rothenberg, S. D. Thompson, I. A. Holm, I. S. Kohane and J. B. Harley (2014). Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. Front Genet 5(401).

O'Malley, K. J., K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle and C. M. Ashton (2005). Measuring Diagnoses: ICD Code Accuracy. Health Serv Res 40(5 Pt 2): 1620-1639.

Peacock, G., D. Amendah, L. Ouyang and S. D. Grosse (2012). Autism spectrum disorders and health care expenditures: the effects of co-occurring conditions. J Dev Behav Pediatr 33(1): 2-8.

Prosser, R. J., B. C. Carleton and M. A. Smith (2008). Identifying Persons with Treated Asthma Using Administrative Data via Latent Class Modelling. Health Serv Res 43(2): 733-754.

Roden, D. M. and J. C. Denny (2016). Integrating electronic health record genotype and phenotype datasets to transform patient care. Clin Pharmacol Ther 99(3): 298-305.

Rutter, M., A. Bailey and C. Lord (2003). SCQ: Social communication questionnaire. Los Angeles, CA, Western Psychological Services.

Schendel, D. E., C. Diguiseppi, L. A. Croen, M. D. Fallin, P. L. Reed, L. A. Schieve, L. D. Wiggins, J. Daniels, J. Grether, S. E. Levy, L. Miller, C. Newschaffer, J. Pinto-Martin, C. Robinson, G. C. Windham, A. Alexander, A. S. Aylsworth, P. Bernal, J. D. Bonner, L. Blaskey, C. Bradley, J. Collins, C. J. Ferretti, H. Farzadegan, E. Giarelli, M. Harvey, S. Hepburn, M. Herr, K. Kaparich, R. Landa, L. C. Lee, B. Levenseller, S. Meyerer, M. H. Rahbar, A. Ratchford, A. Reynolds, S. Rosenberg, J. Rusyniak, S. K. Shapira, K. Smith, M. Souders, P. A. Thompson, L. Young and M. Yeargin-Allsopp (2012). The Study to Explore Early Development (SEED): a multisite epidemiologic study of autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) network. J Autism Dev Disord 42(10): 2121-2140.

Schuler, A., V. Liu, J. Wan, A. Callahan, M. Udell, D. E. Stark and N. H. Shah (2016). Discovering Patient Phenotypes Using Generalized Low Rank Models. Pac Symp Biocomput 21(144-155).

Shahraz, S., T. Lagu, G. A. Ritter, X. Liu and C. Tompkins (2014). Use of Systematic Methods to Improve Disease Identification in Administrative Data: The Case of Severe Sepsis. Med Care

Stacy, M. E., B. Zablotsky, H. A. Yarger, A. Zimmerman, B. Makia and L. C. Lee (2014). Sex differences in co-occurring conditions of children with autism spectrum disorders. Autism 18(8): 965-974.

Verma, A., A. O. Basile, Y. Bradford, H. Kuivaniemi, G. Tromp, D. Carey, G. S. Gerhard, J. E. Crowe, Jr., M. D. Ritchie and S. A. Pendergrass (2016). Phenome-Wide Association Study to Explore Relationships between Immune System Related Genetic Loci and Complex Traits and Diseases. PLoS One 11(8): e0160573.

Wang, L. and D. L. Leslie (2010). Health care expenditures for children with autism spectrum disorders in Medicaid. J Am Acad Child Adolesc Psychiatry 49(11): 1165-1171.

Wang, X., B. Zuckerman, C. Pearson, G. Kaufman, C. Chen, G. Wang, T. Niu, P. H. Wise, H. Bauchner and X. Xu (2002). Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. JAMA 287(2): 195-202.

Zerbo, O., Y. Qian, C. Yoshida, J. K. Grether, J. Van de Water and L. A. Croen (2013). Maternal Infection During Pregnancy and Autism Spectrum Disorders. J Autism Dev Disord.

Zheng, T., W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang and Y. Chen (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 97(120-127).

Chapter 4

Allis, C. D. and T. Jenuwein (2016). The molecular hallmarks of epigenetic control. Nat Rev Genet 17(8): 487-500.

Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30(10): 1363-1369.

Atladottir, H. O., T. B. Henriksen, D. E. Schendel and E. T. Parner (2012). Autism after infection, febrile episodes, and antibiotic use during pregnancy: an exploratory study. Pediatrics 130(6): e1447-1454.

Bakulski, K. M., A. Halladay, V. W. Hu, J. Mill and M. D. Fallin (2016). Epigenetic Research in Neuropsychiatric Disorders: the "Tissue Issue". Curr Behav Neurosci Rep 3(3): 264-274.

Bassil, C. F., Z. Huang and S. K. Murphy (2013). Bisulfite pyrosequencing. Methods Mol Biol 1049(95-107).

Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. B. Fan and R. Shen (2011). High density DNA methylation array with single CpG site resolution. Genomics 98(4): 288-295.

Bibikova, M., J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen and K. L. Gunderson (2009). Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1(1): 177-200.

Blanco-Suarez, E., A. L. Caldwell and N. J. Allen (2017). Role of astrocyte-synapse interactions in CNS disorders. J Physiol 595(6): 1903-1916.

Breton, C. V., H. M. Byun, M. Wenten, F. Pan, A. Yang and F. D. Gilliland (2009). Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. Am J Respir Crit Care Med 180(5): 462-467.

Breton, C. V., Marsit, C. J., Faustman, E., Nadeau, K., Goodrich, J. M., Dolinoy, D. C., … Murphy, S. K. (2017). Small-Magnitude Effect Sizes in Epigenetic End Points are Important in Children's Environmental Health Studies: The Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group. Environmental Health Perspectives, 125(4), 511–526.

Callinan, P.A. and A.P. Feinberg (2006). The emerging science of epigenomics. Hum Mol Genet, 15(1), R95-101.

Casella, G. and R. L. Berger (2002). Statistical Inference, Duxbury.

Cecil, C. A., R. G. Smith, E. Walton, J. Mill, E. J. McCrory and E. Viding (2016). Epigenetic signatures of childhood abuse and neglect: Implications for psychiatric vulnerability. J Psychiatr Res 83(184-194).

Chadwick, L. H., A. Sawa, I. V. Yang, A. Baccarelli, X. O. Breakefield, H.-W. Deng, D. C. Dolinoy, M. D. Fallin, N. T. Holland, E. A. Houseman, S. Lomvardas, M. Rao, J. S. Satterlee, F. L. Tyson, P. Vijayanand and J. M. Greally (2015). New insights and updated guidelines for epigenome-wide association studies. Neuroepigenetics 1(14-19).

Chen, C., K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin and C. Liu (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One 6(2): e17238.

Collier, S. A., S. A. Rasmussen, M. L. Feldkamp, M. A. Honein and S. National Birth Defects Prevention (2009). Prevalence of self-reported infection during pregnancy among control mothers in the National Birth Defects Prevention Study. Birth Defects Res A Clin Mol Teratol 85(3): 193-201.

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414): 57-74.

Davies, M. N., M. Volta, R. Pidsley, K. Lunnon, A. Dixit, S. Lovestone, C. Coarfa, R. A. Harris, A. Milosavljevic, C. Troakes, S. Al-Sarraj, R. Dobson, L. C. Schalkwyk and J. Mill (2012). Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. Genome Biol 13(6): R43.

Du, P., X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou and S. M. Lin (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11(587).

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10): R80.

Hannon, E., K. Lunnon, L. Schalkwyk and J. Mill (2015). Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. Epigenetics 10(11): 1024-1032.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oles, H. Pages, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron and M. Morgan (2015).

Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 12(2): 115-121.

Huen, K., P. Yousefi, A. Bradman, L. Yan, K. G. Harley, K. Kogut, B. Eskenazi and N. Holland (2014). Effects of age, sex, and persistent organic pollutants on DNA methylation in children. Environ Mol Mutagen 55(3): 209-222.

Iurlaro, M., F. von Meyenn and W. Reik (2017). DNA methylation homeostasis in human and mouse development. Curr Opin Genet Dev 43(101-109).

Jaffe, A. E. and R. A. Irizarry (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 15(2): R31.

Jaffe, A. E., P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg and R. A. Irizarry (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol 41(1): 200-209.

Joubert, B. R., S. E. Haberg, R. M. Nilsen, X. Wang, S. E. Vollset, S. K. Murphy, Z. Huang, C. Hoyo, O. Midttun, L. A. Cupul-Uicab, P. M. Ueland, M. C. Wu, W. Nystad, D. A. Bell, S. D. Peddada and S. J. London (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect 120(10): 1425-1431.

Juodzbalys, G., D. Kasradze, M. Cicciu, A. Sudeikis, L. Banys, P. Galindo-Moreno and Z. Guobis (2016). Modern molecular biomarkers of head and neck cancer. Part I. Epigenetic diagnostics and prognostics: Systematic review. Cancer Biomark 17(4): 487-502.

Kurdyukov, S. and M. Bullock (2016). DNA Methylation Analysis: Choosing the Right Method. Biology (Basel) 5(1).

Ladd-Acosta, C. (2015). Epigenetic Signatures as Biomarkers of Exposure. Curr Environ Health Rep 2(2): 117-125.

Ladd-Acosta, C., C. Shu, B. K. Lee, N. Gidaya, A. Singer, L. A. Schieve, D. E. Schendel, N. Jones, J. L. Daniels, G. C. Windham, C. J. Newschaffer, L. A. Croen, A. P. Feinberg and M. Daniele Fallin (2016). Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environ Res 144(Pt A): 139-148.

Lee, K.W., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., Reischl, E., Waldenberger, M., Zeilinger, S., Gaunt, T., McArdle, W., Ring, S., Woodward, G., Bouchard, L., Gaudet, D., Smith, G. D., Relton, C., Paus, T. and Z. Pausova (2015). Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environ Health Perspect. 123(2):193-9.

Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe and J. D. Storey (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28(6): 882-883.

Mehta, D., T. Klengel, K. N. Conneely, A. K. Smith, A. Altmann, T. W. Pace, M. Rex-Haffner, A. Loeschner, M. Gonik, K. B. Mercer, B. Bradley, B. Muller-Myhsok, K. J. Ressler and E. B. Binder (2013). Childhood maltreatment is associated with distinct genomic and epigenetic profiles in posttraumatic stress disorder. Proc Natl Acad Sci U S A 110(20): 8302-8307.

Moran, S., C. Arribas and M. Esteller (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics 8(3): 389-399.

Offenhauser, N., A. Borgonovo, A. Disanza, P. Romano, I. Ponzanelli, G. Iannolo, P. P. Di Fiore and G. Scita (2004). The eps8 family of proteins links growth factor stimulation to actin reorganization generating functional redundancy in the Ras/Rac pathway. Mol Biol Cell 15(1): 91-98.

Olkhov-Mitsel, E. and B. Bapat (2012). Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. Cancer Med 1(2): 237-260.

Rakyan, V. K., T. A. Down, D. J. Balding and S. Beck (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet 12(8): 529-541.

Rutter, M., A. Bailey and C. Lord (2003). SCQ: Social communication questionnaire. Los Angeles, CA, Western Psychological Services.

Schendel, D. E., C. Diguiseppi, L. A. Croen, M. D. Fallin, P. L. Reed, L. A. Schieve, L. D. Wiggins, J. Daniels, J. Grether, S. E. Levy, L. Miller, C. Newschaffer, J. Pinto-Martin, C. Robinson, G. C. Windham, A. Alexander, A. S. Aylsworth, P. Bernal, J. D. Bonner, L. Blaskey, C. Bradley, J. Collins, C. J. Ferretti, H. Farzadegan, E. Giarelli, M. Harvey, S. Hepburn, M. Herr, K. Kaparich, R. Landa, L. C. Lee, B. Levenseller, S. Meyerer, M. H. Rahbar, A. Ratchford, A. Reynolds, S. Rosenberg, J. Rusyniak, S. K. Shapira, K. Smith, M. Souders, P. A. Thompson, L. Young and M. Yeargin-Allsopp (2012). The Study to Explore Early Development (SEED): a multisite epidemiologic study of autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) network. J Autism Dev Disord 42(10): 2121-2140.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3(Article3).

Triche, T. J., Jr., D. J. Weisenberger, D. Van Den Berg, P. W. Laird and K. D. Siegmund (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 41(7): e90.

Um, J. W. (2016). Synaptic functions of the IQSEC family of ADP-ribosylation factor guanine nucleotide exchange factors. Neurosci Res, 116:54-59.

Wiggins, L. D., Levy, S. E., Daniels, J., Schieve, L., Croen, L. A., DiGuiseppi, C., Schendel, D. (2015). Autism Spectrum Disorder Symptoms Among Children Enrolled in the Study to Explore Early Development (SEED). J Autism Dev Disord 45(10), 3183–3194.

Wiggins, L. D., A. Reynolds, C. E. Rice, E. J. Moody, P. Bernal, L. Blaskey, S. A. Rosenberg, L. C. Lee and S. E. Levy (2015). Using standardized diagnostic instruments to classify children with autism in the study to explore early development. J Autism Dev Disord 45(5): 1271-1280.

Zerbo, O., Y. Qian, C. Yoshida, J. K. Grether, J. Van de Water and L. A. Croen (2015). Maternal Infection During Pregnancy and Autism Spectrum Disorders. J Autism Dev Disord 45(12): 4015-4025.

Appendices

Appendix A: Primary Data Collection

1. Experience with data management, quality assurance, and quality control in the Boston Birth Cohort:
    A. extraction of ICD-9-CM codes from electronic medical records, both from children during their pediatric care at the Boston Medical Center, and from the duration of their mother's prenatal care at the BMC
    B. QA/QC in the construction of a nested case-control study from the Children's Health Study pediatric cohort/construction of the analytic dataset

2. Experience with data management, quality assurance, and quality control in a Kabuki Syndrome project with collaborators at the School of Medicine:
    A. QA/QC at the probe- and sample-level for epigenetic data, including reconciliation of predicted vs. recorded patient sex, annotation of mutation status
    B. normalization of data prior to further analysis

3. Experience with data collection, management, quality assurance, and quality control in whole genome sequencing project of a strabismus family with collaborators at the School of Medicine:
    A. study design and data collection (choice of family members based on genetic distance and homogeneity of phenotype)
    B. data management and QA/QC of next generation sequencing data in preparation for analysis

Appendix B: Epigenome-wide Association Study of Kabuki Syndrome

This appendix describes work currently under review at the American Journal of Human Genetics, with contributions from co-authors Nara Sobreira, Li Zhang, Christine Ladd-Acosta, Chrissie Ongaco, Jane Romm, Kimberly F Doheny, Debora Bertola, Chong A Kim, Ana BA Perez, Maria I Melaragno, David Valle, Vera A Meloni, and Hans T Bjornsson.

Kabuki syndrome: New genes and variants with evidence for interplay between histone and DNA methylation machineries

Abstract

Kabuki syndrome (KS) is a monogenic disorder caused by loss of function variants in either of two genes encoding histone-modifying enzymes. We performed targeted sequencing in a cohort of 27 probands with a clinical diagnosis of KS. Of these, 12 had causative variants in the two known KS genes. In 2, we identified presumptive loss of function *de novo* missense variants in *KMT2A*, a gene that encodes another histone modifying enzyme previously exclusively associated with Wiedermann-Steiner syndrome. Surprisingly, we also find alterations in DNA methylation among individuals with a KS diagnosis relative to matched normal controls regardless of whether they carry variant in *KMT2A* or *KMT2D* or not. Furthermore, we observed characteristic global abnormalities of DNA methylation that distinguished patients with a loss of function

variant in *KMT2D* or missense changes in either *KMT2D* or *KMT2A* from normal controls. Our results provide new insights into the relationship of genotype to epigenotype and phenotype and indicate cross-talk between histone and DNA methylation machineries exposed by inborn errors of the epigenetic apparatus.

Kabuki syndrome (KS; MIM 147920, 300867) is a pleiotropic disorder characterized by intellectual disability, postnatal growth retardation and dysmorphic facial features[1,2]. Defects of B cell differentiation are also frequent[3]. In about 75% of KS individuals, the disorder is inherited as an autosomal dominant trait caused by loss of function (LOF) variants in *KMT2D*[1] (previously known as *MLL2*); in another 6% of the cases, the disorder shows X-linked dominant inheritance caused by LOF variants in *KDM6A*[2]. The explanation for the remaining ~20% of KS cases is unknown.

KMT2D encodes lysine-specific histone methyltransferase (KMT2D) that catalyzes methylation of H3K4. *KDM6A* encodes a lysine-specific demethylase (KDM6A) that catalyzes removal of methyl groups from H3K27me3. Thus, both KS-associated genes regulate histone tail methylation. Little is known about the target genes for KMT2D and KDM6A in normal human cells and their relevance to the KS phenotype[4].

Wiedemann-Steiner syndrome (WSS; MIM 605130) is a rare autosomal dominant disorder caused by heterozygous loss of function variants in the *KMT2A* gene[5]. Jones et

al. identified heterozygous, *de novo* nonsense or indel variants in *KMT2A* in 5 out of 6 individuals with a phenotype characterized by hypertrichosis cubiti, excessive hair on the back with a whorl-like distribution, long eyelashes, thick or arched eyebrows with a lateral flare, and down-slanting and vertically narrow palpebral fissures, sacral dimple, height bellow the 10th centile, mild to moderate intellectual disability and behavioral difficulties[5].

Here we studied a large KS cohort to expand our knowledge of genotype-phenotype relationships in KS. We performed targeted sequencing and genome-wide epigenomic analyses to: (1) identify novel genetic variants associated with KS, and (2) characterize KS-associated epigenomic patterns, generally, as well as the relationship between specific genetic variants identified in KS patients and epigenomic profiles. We examined 27 probands with a KS phenotype seen at Escola Paulista de Medicina, Sao Paulo and at Universidade de Sao Paulo, Brazil submitted to the Baylor-Hopkins Center for Mendelian Genomics (BHCMG) through the online submission portal PhenoDB[6] as well as 9 samples from control individuals that were matched on age, sex and ethnicity. The clinical diagnosis was based on the presence of the most common features seen in KS[7-8]. We performed targeted next-generation sequencing of 9 genes (Table B.1), including the 2 genes known to cause KS (*KMT2D* and *KDM6A*, Table B.2); 2 genes known to cause ICF syndrome (*DNMT3B* and *ZBTB24*), a recognized genocopy of KS (Table B.2); and 5 candidate genes (*KDM6B*, *MEN1*, *KMT2A*, *KMT2B*, *HCFC1*) known to interact with or have overlapping function with known KS genes.

Table B.1: Genes selected for targeted sequencing

Gene Symbol	Reason for selection
KMT2D (MLL2)	Known KS gene
KDM6A	Known KS gene
DNMT3B	ICF gene (KS genocopy)
ZBTB24	ICF gene (KS genocopy)
KDM6B	Candidate gene
MEN1	Candidate gene
KMT2A	Candidate gene
KMT2B	Candidate gene
HCFC1	Candidate gene

ICF, Immunodeficiency-centromeric instability-facial anomalies syndrome; **KS,** Kabuki syndrome

Table B.2: Discussed disorders, Features, Genes, Function and Epigenetic Consequences

	Features	Genes	Function	Epigenetic effect	MIM#	Primary defect
Kabuki syndrome	Intellectual disability, dysmorphic face, growth retardation, immune dysfunction	*KMT2D* *KDM6A*	Histone methyltransferase Histone demethylase	↓H4K4me-me3 ↑H3K27me3	147920 300867	Histone
Wiedemann-Steiner syndrome	Intellectual disability, dysmorphic face, hirsutism	*KMT2A*	Histone methyltransferase	↓H4K4me-me3	605130	Histone
Sotos syndrome	Learning disability, overgrowth, large forehead	*NSD1* *NFIX*	Histone methyltransferase Transcription factor	↓H3K36me, ↓H4K20me Unknown	117550 614753	Histone
ICF syndrome	Intellectual disability, dysmorphic face, growth retardation, immune dysfunction	*DNMT3B* *ZBTB24* *CDCA7* *HELLS*	*De novo* DNA methyltransferase Zinc finger transcription repressor Transcriptional regulator Chromatin remodeler	↓DNA methylation ↓DNA methylation ↓DNA methylation ↓DNA methylation	242860 614069 616910 616911	DNA methylation
Mental retardation, X-linked 3	Intellectual disability, methylmalonic aciduria, homocysteinemia	*HCFC1*	Transcription factor that potentially recruits histone machinery	Unknown	309541	Unknown

We designed probes for targeted sequencing (TruSeq Custom Amplicon kit, Illumina) with the online Illumina DesignStudio software for all candidate genes including exon-intron boundaries. We prepared the sequencing library according to the manufacturer´s protocol and sequenced one library pool of 28 samples (27 individuals and 1 control) in a single run on a MiSeq sequencer (Illumina, 2 X 251 bp paired end reads). Alignment of NGS data to the human reference genome and variant calling were performed using software provided by Illumina. All variant calls were based on RefSeq transcript and NCBI human genome assembly build 37. We then used the Analysis Tool of PhenoDB[9] to prioritize rare heterozygous and homozygous functional variants (missense, nonsense, splice site variants and indels) and excluded variants with a MAF > 0.01 in the Exome Variant Server (release ESP6500SI-V2) or 1000 Genomes Project[10] and variants present in dbSNP 126, 129, or 131. Next we generated a heterozygous, homozygous and a compound heterozygous variant list for each subject and every candidate single nucleotide variant and indel were verified by inspection in Integrative Genomics Viewer[11].

In 12 KS probands, we identified *KMT2D* variants: 3 with unique *de novo* heterozygous variants (Table B.3). Eight of the 12 *KMT2D* variants we identified are not present in the Exome Aggregation Consortium (ExAC) database, three variants are present as heterozygous variants in one individual each and one is present as a heterozygous variant in three individuals. In a fourth individual (KS11) we found a missense variant in *KMT2D* (p.N4572S) and a missense variant in *KMT2B* (p.E2354K). The *KMT2D* variant is rare (0.0008% in ExAc), does not involve a known protein domain

and has not been previously associated with KS, while the *KMT2B* variant is novel and alters a conserved residue (E2354). KMT2B is a histone methyltransferase that also targets H3K4, raising the possibility that in KS11 either or both of these variants contribute to the KS phenotype. The phenotype of individual KS11 does not seem to be atypical and fits the clinical diagnosis of KS.

Table B.3: Summary of variants identified with targeted next-generation sequencing among 26 individuals with clinically defined Kabuki syndrome

Subject	Gene	Variant			Parents sequenced	ExAC MAF	DNA methyl-ation[a]
		Nucleotide change	Amino acid change	Exon			
Known KS genes							
KS4	*KMT2D*	c.14515+1G>T		47	N	0	+
KS7	*KMT2D*	c.C12268T	p.Q4090X	39	N	0	+
KS9	*KMT2D*	c.5124_5125del	p.T1708fs	21	N	0	+
KS11	*KMT2D*	c.A13715G	p.N4572S	41	N	0.0008%	−
KS12	*KMT2D*	11582_11583insGC	p.Q3861delinsQQ	39	N	0.005%	+
KS13	*KMT2D*	c.C185T	p.P62L	3	Mother (not present)	0.002%	+
KS14	*KMT2D*	c.1329_1332del	p.P443fs	10	N	0.0009%	+
KS15	*KMT2D*	c.6595delT	p.Y2199fs	31	Y (*de novo*)	0	+
KS19	*KMT2D*	c.15920_15921insT	p.S5307fs	49	Mother (not present)	0	+
KS21	*KMT2D*	c.13207_13208del	p.N4403fs	39	Y (*de novo*)	0	+
KS22	*KMT2D*	c.C12304T	p.Q4102X	39	Y (*de novo*)	0	+
KS24	*KMT2D*	c.C11800T	p.Q3934X	39	N	0	+
Other candidate genes sequenced							
KS8	*KMT2A*	c.G3019T	p.G1007C	3	Y (*de novo*)	0	+

KS29	*KMT2A*	c.5803-1G>A		22	Y (*de novo*)	0	+
KS5	*HCFC1*	c.C3795T	p.S1265L	17	N	0.07%	−
KS10	*ZBTB24*	c.G146A	p.R49Q	2	Mother (not present)	0.6%	−
KS11	*KMT2B*	c.G7060A	p.E2354K	30	N	0	−
KS18	*DNMT3B*	c.A1211G	p.Y404C	11	N	0.8%	−

KS, Kabuki Syndrome; *KMT2A*, lysine (K)-specific methyltransferase 2A; *KMT2D*, lysine (K)-specific methyltransferase 2D; *HCFC1*, host cell factor C1; *DNMT3B,* DNA methyltransferase 3 beta; *ZBTB24*, zinc finger and BTB domain containing.

[a]KS11 was removed from the methylation analysis because it had variants in both *KMT2B* and *KMT2D,* and was unique in that the *KMT2D* variant was a missense variant of unknown functional consequence.

[b]For KS1-3, 6, 16-17, 20, 26-28 no mutation was found.

[c]Plus and minus signs indicate a sample was included or excluded from DNA methylation analyses, respectively. *KMT2D* transcript identifier is NM_003482, *KMT2A* transcript identifier is NM_001197104, *HCFC1* transcript identifier is NM_005334, *ZBTB24* transcript identifier is NM_001164313, *KMT2B* transcript identifier is NM_014727, *DNMT3B* transcript identifier is NM_006892.

[d]In one female individual (KS5) we identified a heterozygous variant in *HCFC1* (MIM 309541), a gene known to be responsible for an X-linked disorder phenotypically distinct from KS (Table S2).

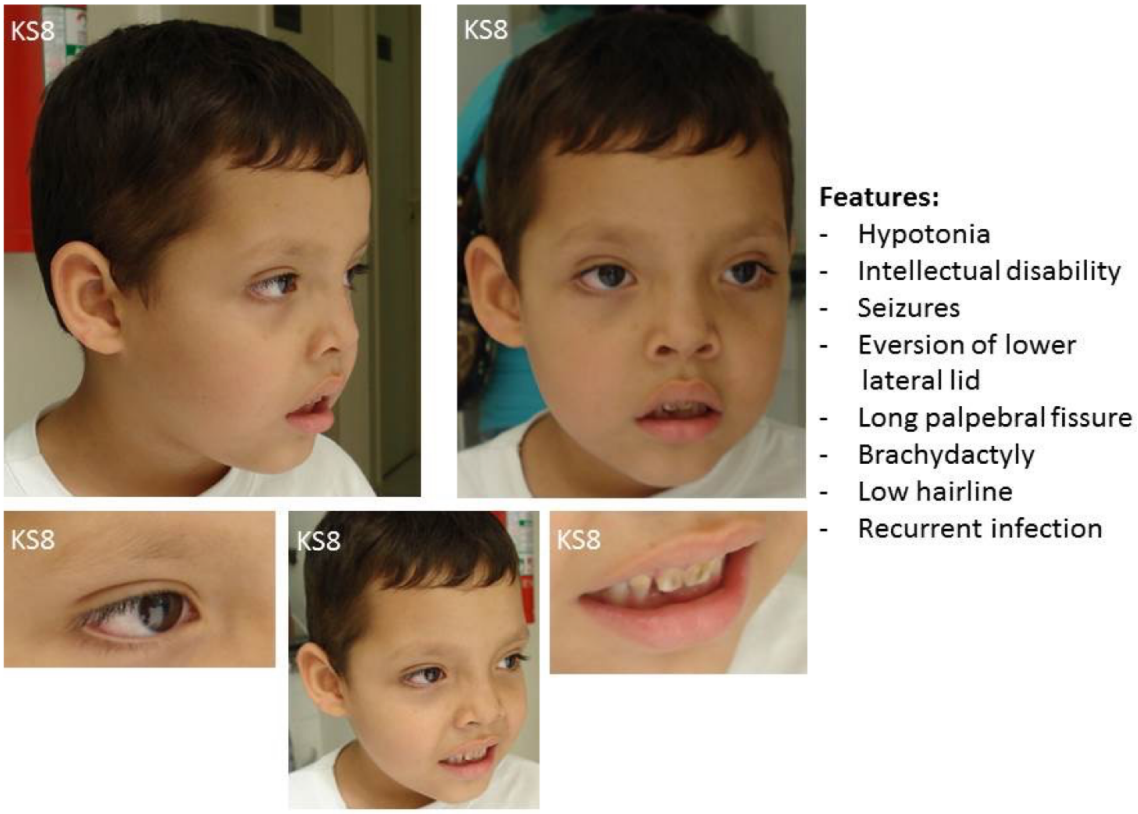
We also found two individuals with a typical KS phenotype (KS8 and 29) who had normal *KMT2D* and *KDM6A* sequence but had novel *de novo* heterozygous missense variants in *KMT2A*, a histone methyltransferase gene previously associated with Wiedemann-Steiner syndrome (WSS). WSS is a disorder with overlapping features with KS[5] such as vertically narrow palpebral fissures, strabismus, broad nasal bridge/tip, external ear deformity, short stature, hypotonia, small hands, hip abnormalities, developmental delay, intellectual disabilities, seizures, feeding difficulties, hypertrichosis, heart anomalies, urological anomalies, and recurrent infections (Table B.2). Review of the phenotype of the two individuals with the *KMT2A* variants confirmed the clinical diagnosis of KS (Figure B.1a and B.1b). The comparison of their facial features to the patients diagnosed with WSS revealed significant facial similarities with the patients

described by Jones et al. (WSS-5)[5], Mendelssohn et al. (2014)[12] and Stellacci et al. (2016)[13]. These results suggest a significant phenotypic overlap between KS and WSS suggesting that pathological variants in *KMT2A* should be considered in KS individuals who lack variants in *KMT2D* or *KDM6A*. Interestingly, the proteins encoded by the *KMT2A* and *KMT2D* (WSS and KS, respectively) genes have many of the same domains (Figure B.1c). We also identified an individual (KS18) with a heterozygous variant in *DNMT3B* and one individual (KS10) with a heterozygous variant in *ZBTB24*. Both of these genes have previously been associated with ICF syndrome, an autosomal recessive known genocopy of KS (Table B.2). These patients have a typical KS phenotype without any atypical features or history of recurrent infections. Individual KS7 had a *KMT2D* p.Q4090X variant and an atypical feature, hypoplasia/aplasia of the medial and/or distal phalanges of the toes bilaterally. Similarly, individual KS15, with the *KMT2D* p.Y2199fs variant, had shortening of the medial phalange of the 2nd and 5th fingers bilaterally and syndactyly of the 4th and 5th toes bilaterally. Skeletal anomalies are characteristics of KS but they most commonly affect the hands, not the feet and toes. In thirteen KS probands (48%) we failed to identify likely pathogenic variants in any of the 9 genes. This negative result is similar to that (36%) of a recent large sequencing study of patients with KS[12]. Direct Sanger sequencing of PCR amplified products validated all variants and confirmed appropriate Mendelian segregation in available family members (Table B.3).

Figure B.1: Shared facial features in patients with variants in *KMT2A* and *KMT2D*

a.

b.



c.

Shared facial features in patients with variants in two Trithorax orthologs (*KMT2A* and *KMT2D*). Facial features (left) in two patients identified as KS29 (A) and KS8 (B) led to a clinical diagnosis of Kabuki syndrome. Here we have summarized the observed phenotype (right). These two patients suggest that there is phenotypic overlap between Kabuki and Wiedemann-Steiner syndromes, two disorders that are caused by variants in two independent Trithorax orthologs (*KMT2A* and *KMT2D*). The domain structure of the two orthologous proteins (C). Domain structure is based on InterPro.

To further elucidate potential mechanisms of the primary histone alterations occurring in KS, we measured DNA methylation levels across the genome in a total of 44 samples including 27 KS and 9 age- and sex-matched control samples. 1125 ng of genomic DNA was prepared, for each sample, in a total volume of 45µl and bisulfite treated using the EZ DNA Methylation kit (Zymo Research Corp, Orange, CA, USA), as specified by the manufacturer for use with 450K arrays. To obtain genome-scale methylation measurements, bisulfite treated DNA was processed on the Infinium

HumanMethylation450 BeadChip (Illumina, San Diego, CA, USA) at the Johns Hopkins SNP Center in accordance with the manufacturer's recommendation. The Infinium HumanMethylation450 BeadChip measures DNA methylation levels at 485,512 loci across the genome. A total of 44 samples, including 27 KS samples and 9 non-KS controls were randomized across (4) 12 array BeadChips to minimize potential confounding batch effects. For quality control purposes we also included technical control samples with known amounts of DNA methylation (0%, 25%, 50%, 75%, and 100% methylated). All data preprocessing steps and statistical analyses were performed using R-3.1.2 and Bioconductor 3.0[15]. Using the 'minfi' package[16], we applied several quality control measures including removal of samples with low overall array intensities (n=0), as well as removal of poorly performing probes (n=2,633), defined as having a detection p value greater than 0.01 in any sample. Two of the samples had a predicted sex that was discordant with the sample annotation, which was reconciled with clinical notes before proceeding. After dropping parental controls (n=6) due to the known epigenetic changes that occur with aging, thirty-eight samples, including 9 age- and sex-matched controls, and 2 samples who were later removed from further analysis due to a change in their KS diagnosis, were retained for calculation of principle components for ancestry adjustment[20]. All samples then underwent quantile normalization and beta values were logit-transformed to create M-values, which are normally distributed and therefore more appropriate for statistical testing[17]. Because the DNA methylation measurements were obtained from whole blood samples with a heterogeneous mix of nucleated white cells, we estimated the proportion of B cells, CD4 and CD8 positive T cells, natural killer cells,

monocytes, and granulocytes[18]. Although we did not find significant differences ($p<0.05$) in cell composition between KS and control individuals, we adjusted for cell composition estimates in our analyses to identify KS-associated methylation changes.

Initial genome-wide analyses comparing all KS samples to controls revealed differences in DNA methylation that appeared to be mainly driven by samples with a detected mutation in a histone machinery gene (*KMT2A*/*KMT2D*) (Figure B.2). Based on this observation, the fact that we had a relatively large number of samples with a mutation in a histone machinery genes, and because we were interested in identifying downstream epigenetic changes related to mutations in chromatin machinery genes, we decided to focus our primary genome-wide screening analyses on identifying differences in DNA methylation associated with KS among individuals with a molecularly confirmed variant in the histone machinery. Therefore, we removed 16 samples from our DNA methylation analyses that did not have a variant in the histone methylation machinery genes we examined as well as 1 sample that had a variant in more than 1 gene, leaving us with 13 KS cases and 9 controls (n=22). This left a total of 13 blood samples obtained from individuals with a clinical diagnosis of KS and defects in the histone machinery (KMT2A and KMT2D variants) as compared to 9 age and sex matched control individuals (Figure B.3). Differentially methylated positions (DMPs) were identified using limma[19] and linear regression models were adjusted for sex, blood cell composition, and ancestral population[20]. Differentially methylated regions (DMRs) were identified using bumphunter[21]; the analytic model was adjusted for sex, blood cell composition, and ancestral population[20]. DMR significance was assessed using

bootstrapping as available in minfi[16] and a family-wide error rate threshold of 0.05 was applied.

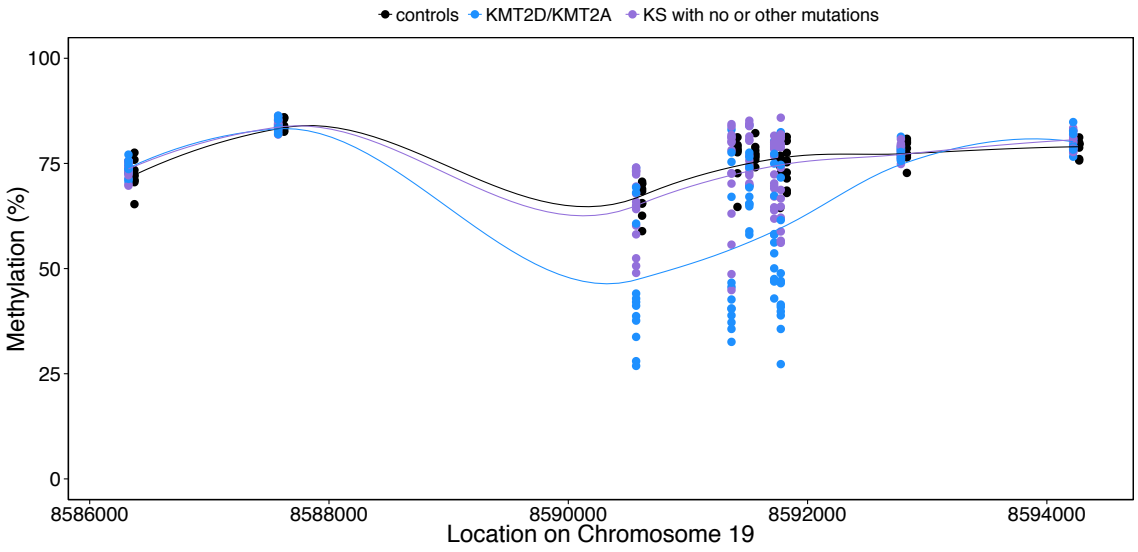


Figure B.2: An example of a differentially methylated region (DMR) identified by comparing all KS samples to normal controls.

Differences appear to be driven by samples with a *KMT2D*/*KMT2A* variant. We therefore decided to focus on molecularly confirmed KS samples for primary genome-wide screening purposes. Genomic location is plotted on the x-axis and percent methylation on the Y-axis. Blue, purple, and black points denote individuals harboring a *KMT2D*/*KMT2A* variant, KS phenotype but no detected variant, and controls, respectively.

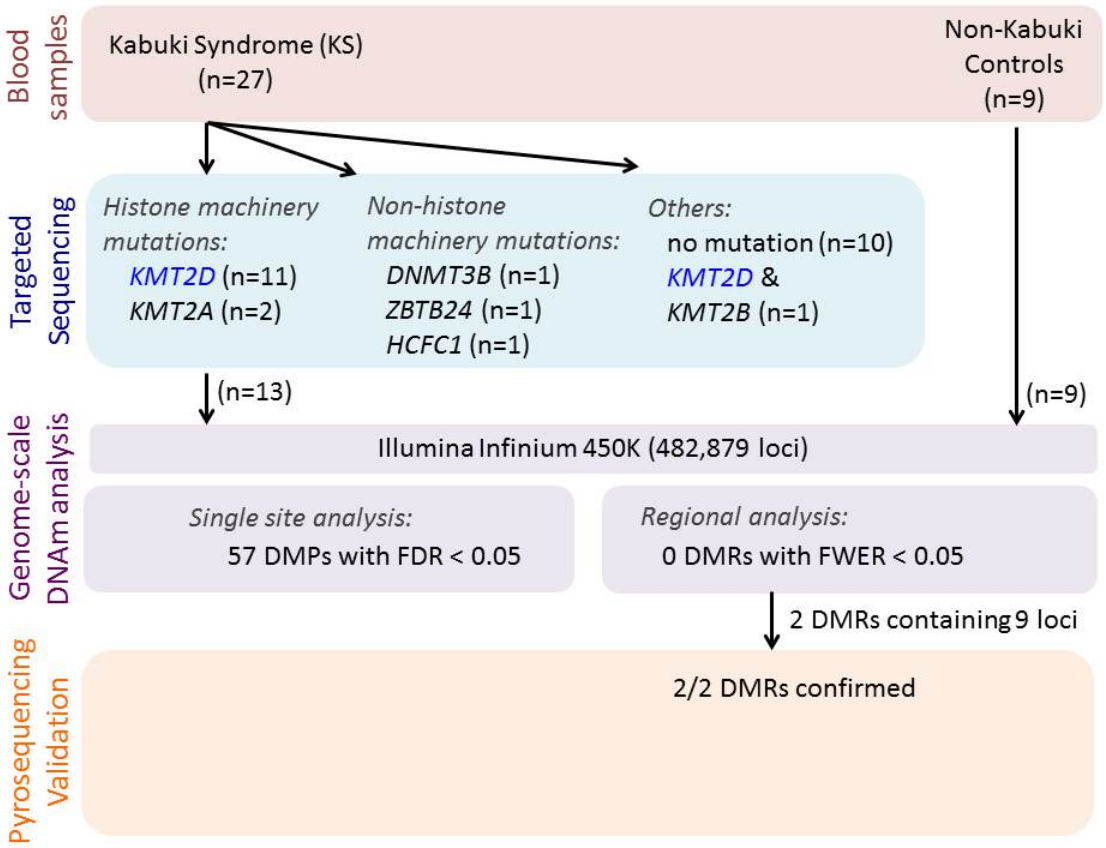


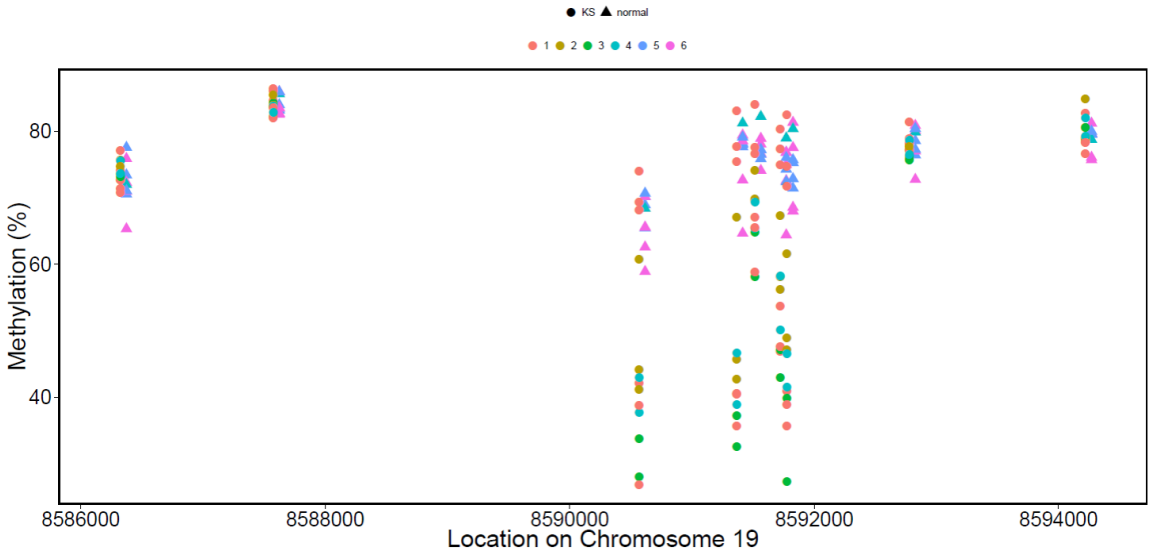
Figure B.3: A flow chart describing sample numbers analyzed in each stage of our analysis.

Since position on array is a potential source of bias we explored the location of our samples. We noted that many of the control samples were processed in rows 5 and 6 and many of the KS samples were processed in rows 1-4 of the Illumina BeadChip. To confirm that the DNA methylation changes related to KS were not solely attributable to array position, we plotted DNA methylation values and array row position for the DMRs reported here (Figure B.4). We observed overlapping DNA methylation levels among samples processed in rows 1-6; thus, DNA methylation values at these KS-related sites are not purely driven by row alone. Furthermore, our findings for these regions were

validated using bisulfite pyrosequencing, an independent and highly reliable method, where array row and location is irrelevant.

When we compared DNA methylation in 13 KS probands with 9 age and sex-matched controls (Figure B.3) we found 57 differentially methylated loci, at an FDR<0.05, associated with KS (Table B.4). Two CpG sites in two independent genes (*CIRBP*, *FEM1B*) show relative hypomethylation among individuals with KS and histone machinery variants compared to controls (Figure B.5a-b). Similarly, Figure B.5c-d shows an example of two genes (*c10orf11*, *SOX18*) that are hypermethylated in KS relative to controls. We also used bumphunter[21], as an alternative analysis to identify differentially methylated regions (DMRs) and found several genomic regions showing KS-related differences in methylation (Table B.5). Although none of these regions reached genome-wide statistical significance in our relatively small sample, several showed striking differences in methylation between KS and controls. Figure B.6a provides an example of a hypomethylated DMR (28% less methylation in KS, on average) in *MYO1F*.

a.



b.

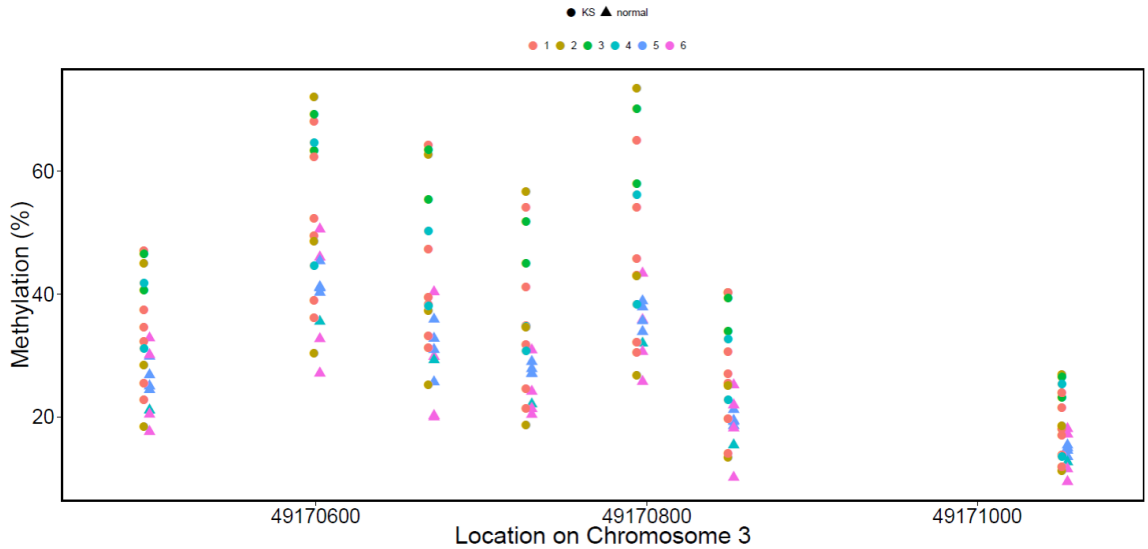


Figure B.4: Plots showing the relationship between DNA methylation level and sample row at two KS-associated DMRs, *MYO1F* DMR (A) and *LAMB2* DMR (B).

KS and control samples are denoted by circles and triangles, respectively. Array row location is denoted by color with row 1 shown in red, row 2 shown in orange, row 3 shown in green, row 4 shown in aqua, row 5 shown in blue, and row 6 shown in purple.

Table B.4: Differentially methylated positions (DMPs) significantly associated (FDR<0.05) with KS patients harboring variants in histone methylation machinery genes compared to non-KS controls

Chromosome	Position	Gene	ΔM[a]	Adjusted *P* value	Adjusted q-value	Annotated SNP[b]
chr19	1272853	*CIRBP*	-7.7%	6.70E-08	0.0167	
chr15	68571585	*FEM1B*	-6.2%	6.90E-08	0.0167	
chr19	45720949	*EXOC3L2*	-15.0%	1.62E-07	0.0186	
chr2	114196091	*CBWD2*	-3.6%	1.91E-07	0.0186	
chr16	88871329	*CDT1*	-7.7%	1.93E-07	0.0186	
chr6	15506085	*JARID2*	-10.2%	4.14E-07	0.0317	
chr10	77871958	*C10orf11*	6.5%	5.09E-07	0.0317	
chr20	62681428	*SOX18*	5.4%	5.25E-07	0.0317	
chr19	45737623	*EXOC3L2*	-16.7%	8.29E-07	0.0332	
chr10	77872084	*C10orf11*	9.1%	1.02E-06	0.0332	
chr17	80202961	*CSNK1D*	-10.8%	1.03E-06	0.0332	
chr4	145655974	*HHIP*	8.4%	1.06E-06	0.0332	
chr15	30206862		7.2%	1.11E-06	0.0332	
chr20	62681243	*SOX18*	15.2%	1.12E-06	0.0332	
chr1	152595992	*LCE3A*	6.7%	1.19E-06	0.0332	
chr1	10699604	*CASZ1*	5.8%	1.24E-06	0.0332	
chr11	859670	*TSPAN4*	-9.0%	1.42E-06	0.0332	
chr22	45094531	*PRR5*	11.9%	1.50E-06	0.0332	
chr7	55086890	*EGFR*	1.9%	1.51E-06	0.0332	
chr17	73630199	*RECQL5*	4.2%	1.52E-06	0.0332	rs185961263
chr3	45077254	*CLEC3B*	6.7%	1.57E-06	0.0332	
chr19	51171712	*SHANK1*	-4.3%	1.58E-06	0.0332	
chr19	1873591		-9.9%	1.58E-06	0.0332	
chr17	79816559	*P4HB*	-10.7%	2.04E-06	0.0402	

Table B.4 continued

chr3	96168838		6.1%	2.12E-06	0.0402	
chr15	47477911		9.5%	2.16E-06	0.0402	
chr18	60264554		-5.2%	2.37E-06	0.0412	
chr19	45720771	*EXOC3L2*	-9.5%	2.40E-06	0.0412	
chr16	57918043	*CNGB1*	5.7%	2.47E-06	0.0412	
chr4	5708474	*EVC2*	-10.0%	2.68E-06	0.0419	
chr12	50017412	*PRPF40B*	-4.5%	2.69E-06	0.0419	
chr2	20442088		4.9%	2.84E-06	0.0428	
chr9	132386878		2.5%	3.02E-06	0.0437	
chr15	30216986		11.8%	3.07E-06	0.0437	
chr14	91862864	*CCDC88C*	6.9%	3.27E-06	0.0437	
chr15	30175531		9.1%	3.30E-06	0.0437	
chr9	44420179		4.1%	3.56E-06	0.0437	
chr11	98940816	*CNTN5*	-6.2%	3.69E-06	0.0437	
chr2	105276153		6.6%	3.72E-06	0.0437	
chr16	88152486		11.7%	3.74E-06	0.0437	
chr2	8978042	*KIDINS220*	5.8%	3.75E-06	0.0437	
chr7	4778881	*FOXK1*	6.0%	3.84E-06	0.0437	
chr10	77871618	*C10orf11*	10.2%	3.89E-06	0.0437	
chr11	850296	*TSPAN4*	-12.8%	4.11E-06	0.0451	
chr19	18260515	*MAST3*	-8.8%	4.36E-06	0.0461	
chr17	79426432	*BAHCC1*	-4.8%	4.47E-06	0.0461	
chr11	910094	*CHID1*	-3.7%	4.49E-06	0.0461	
chr11	71276654	*KRTAP5-10*	-8.4%	4.76E-06	0.0472	rs188029416
chr11	89169539	*NOX4*	5.6%	4.79E-06	0.0472	rs76916726
chr19	846179	*PRTN3*	-4.6%	5.18E-06	0.0486	
chr5	176790179	*RGS14*	9.7%	5.19E-06	0.0486	

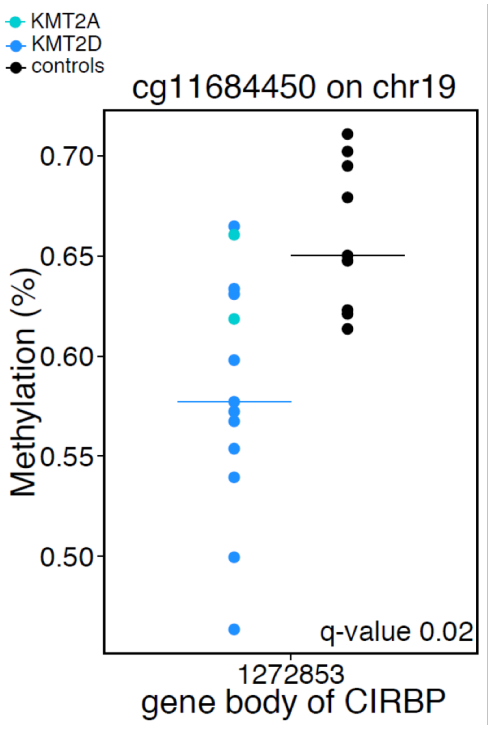
Table B.4 continued					
chr1	147253401		6.4%	5.23E-06	0.0486
chr5	173021074		-5.2%	5.59E-06	0.0499
chr2	132152878		5.8%	5.65E-06	0.0499
chr6	100882035	*SIM1*	6.7%	5.72E-06	0.0499
chr19	815090	*LPPR3*	-11.1%	5.81E-06	0.0499
chr22	19710163	*GP1BB; SEPT5*	10.3%	5.89E-06	0.0499

[a] Difference in mean methylation levels between patients with KS and variants in histone machinery genes and non-KS control groups. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals with KS compared to controls.

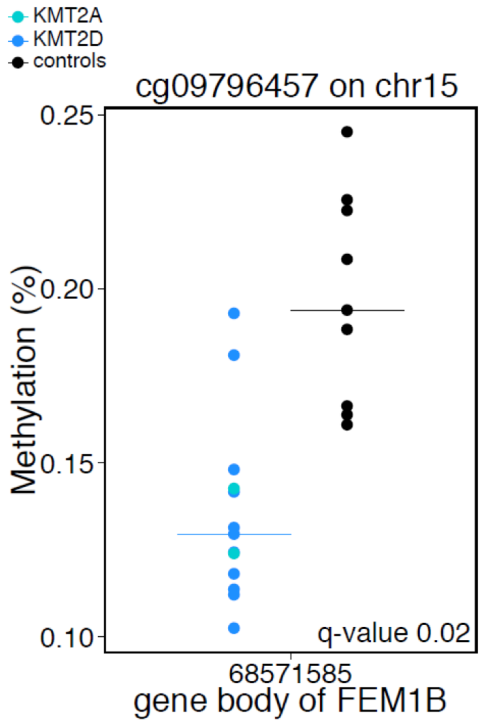
[b] Denotes which probes have an annotated SNP at the measured CpG site and provides the SNP identifier. SNP annotation information was based on dbSNP137 and was obtained using the getAnnotation() function from the minfi[14] Bioconductor package.

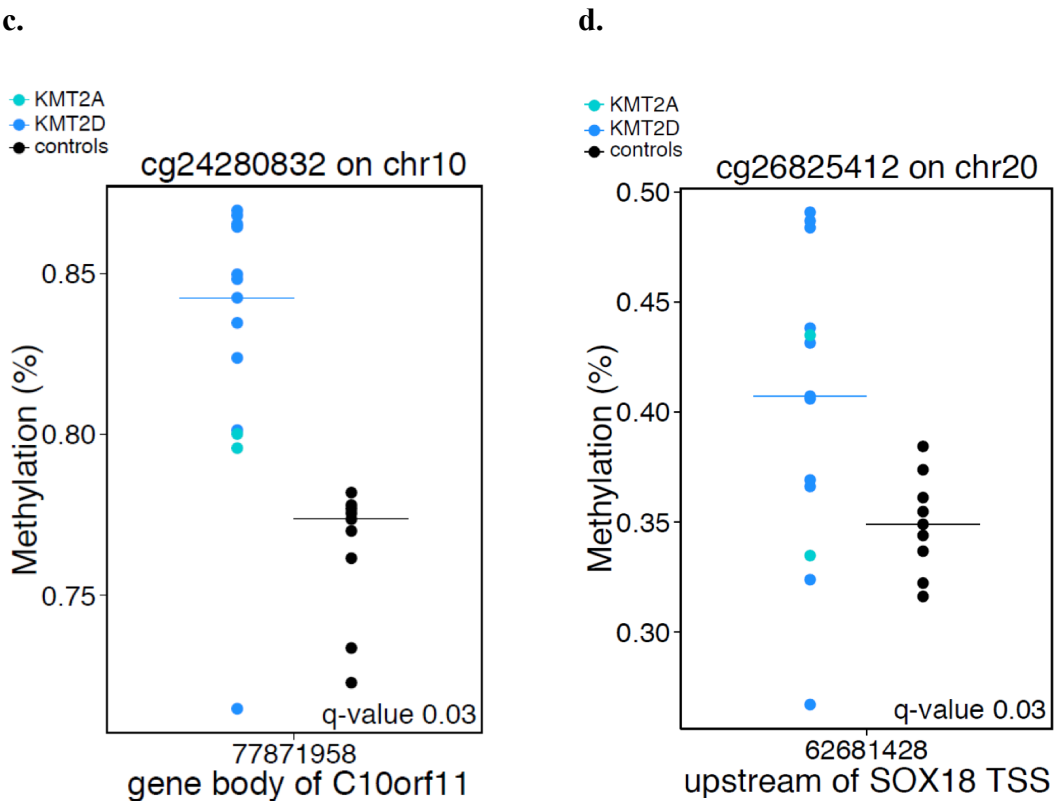
Figure B.5: Differentially methylated positions in individuals with a Kabuki syndrome phenotype

a.



b.





Individuals with Kabuki syndrome phenotype and a variant in a histone methylation machinery gene, either *KMT2D* or *KMT2A*, show site-specific hyper- and hypo-methylation relative to non-KS controls. Genomic loci showing significant relative loss of methylation (*CIRBP*, *FEM1B*) in KS samples compared to controls (a, b). Genomic loci showing significant gain of methylation (*c10orf11, SOX18*) in KS samples compared to controls (c, d). Blue and black points denote KS and non-KS controls, respectively. Light green points denote the two individuals with *KMT2A* variants.

Table B.5: Top 10 differentially methylated regions (DMRs) associated with individuals with KS and variants in a histone methylation machinery gene compared to non-KS controls.

Chr	Position	Nearest gene	DMR location	ΔM[a]	Nominal *P*	FWER[b]
chr5	135415948-135416613	*VTRNA2-1*	Overlaps	-18.6%	1.47E-05	0.203
chr19	8591364-8591776	*MYO1F*[c]	Inside exon	27.6%	7.65E-05	0.813
chr3	49170496-49171051	*LAMB2*[c]	5' UTR	20.0%	9.33E-05	0.627
chr4	74847646-74848016	*PF4*	5' UTR	-19.0%	1.11E-04	0.549
chr5	1594282-1594733	*SDHAP3*	5' UTR	19.4%	1.12E-04	0.654
chr7	27170241-27170755	*HOXA4*	5' UTR	16.2%	1.59E-04	0.594
chr8	1365049-1365749	*DLGAP2*	83.8Kb upstream	-18.6%	1.98E-04	0.809
chr7	27183196-27183686	*HOXA5*	5' UTR	-12.0%	2.57E-04	0.755
chrX	8751190-8751524	*FAM9A*	17.9Kb downstream	20.4%	2.61E-04	0.856
chr6	30039374-30039442	*RNF39*	Inside exon	14.0%	4.04E-04	0.806

Abbreviations: FWER, family-wise error rate

[a] Difference in mean methylation levels between individuals with KS and histone machinery variant and non-KS control groups. Positive values reflect relative hypermethylation and negative values relative hypomethylation in individuals with KS compared to controls.

[b] Empirical significance value, computed using permutation testing (n=1000).

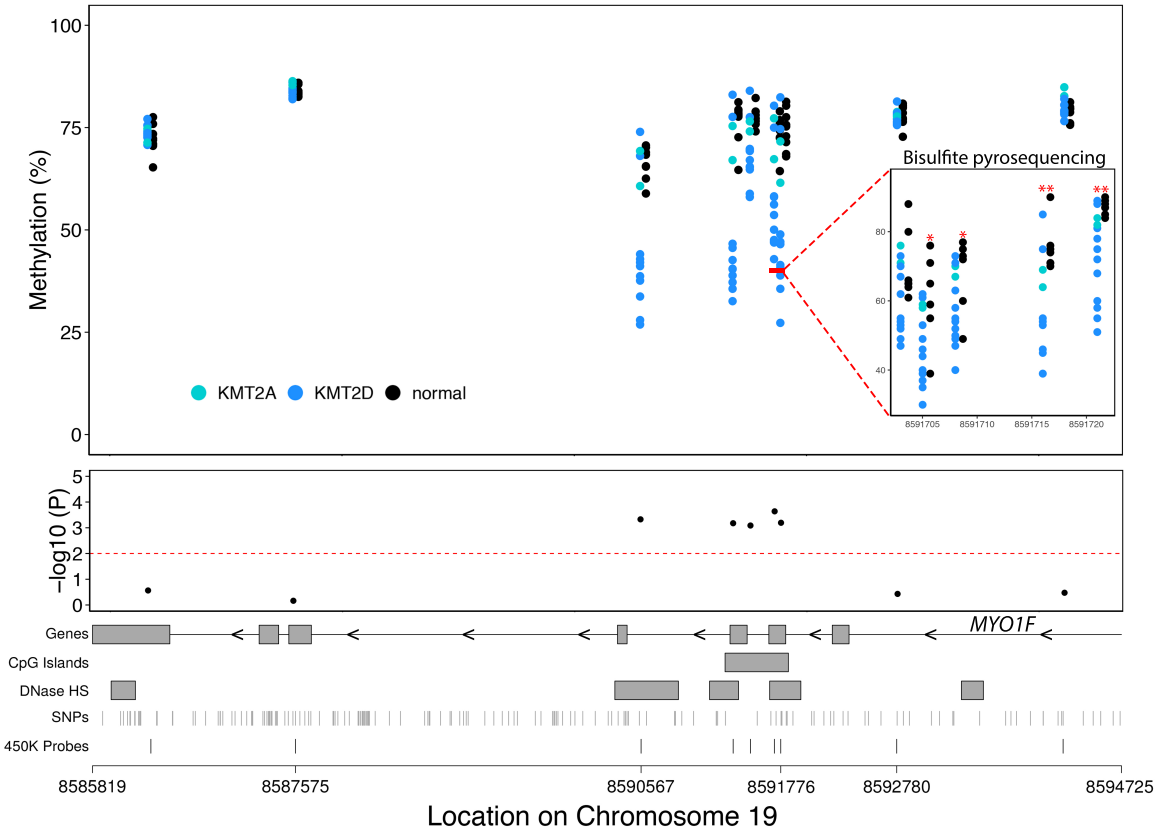
[c] Validated using bisulfite pyrosequencing

To validate these results with an independent assay and to determine whether similar changes in DNA methylation exist in KS samples without a known histone machinery variant we performed bisulfite pyrosequencing for 36 individuals including 12 with *KMT2D* or *KMT2A* variants and KS diagnosis and 14 without *KMT2D*/*KMT2A* variants with KS diagnosis and 10 controls. PCR Primers and sequencing primers were designed by MethPrimer[22] and are available by request. We bisulfite treated all samples used for the Genome Scale DNA methylation measurement and available parents (Controls = 16, no *KMT2D* mutation = 17, *KMT2D*/*KMT2A* mutation = 13). After bisulfite treatment, only a subset of these samples had the minimal amount needed for PCR-based studies (Controls = 7, no *KMT2D*/*KMT2A* mutation = 12, and *KMT2D*/*KMT2A* mutation = 13) and these were used for the validation studies. Bisulfite treated genomic DNA[23] was PCR amplified (50-cycles) in a total of 25ul volume and 5ul were loaded onto agarose gel to verify single, strong product and an absence of any unused primers. The biotinylated PCR product was captured on streptavidin-coated sepharose beads (GE Healthcare, Milwaukee, WI) and pyrosequencing reaction were set up using the PyroMark Gold Q24 kit (Qiagen), according to the manufacturer's instructions. Each individual pyrosequencing assay was designed with PyroMark Q24 software and the percentage of methylation at each CpG site was analyzed with PyroMark Q24 software. We set the p-value threshold for significance to <0.05. Consistent with our array-based findings, KS individuals with a *KMT2D* or *KMT2A* variant show relative hypomethylation at the *MYO1F* locus (Figure B.6b), however, we also noted that patients with a KS phenotype but without molecular confirmation also demonstrated relative

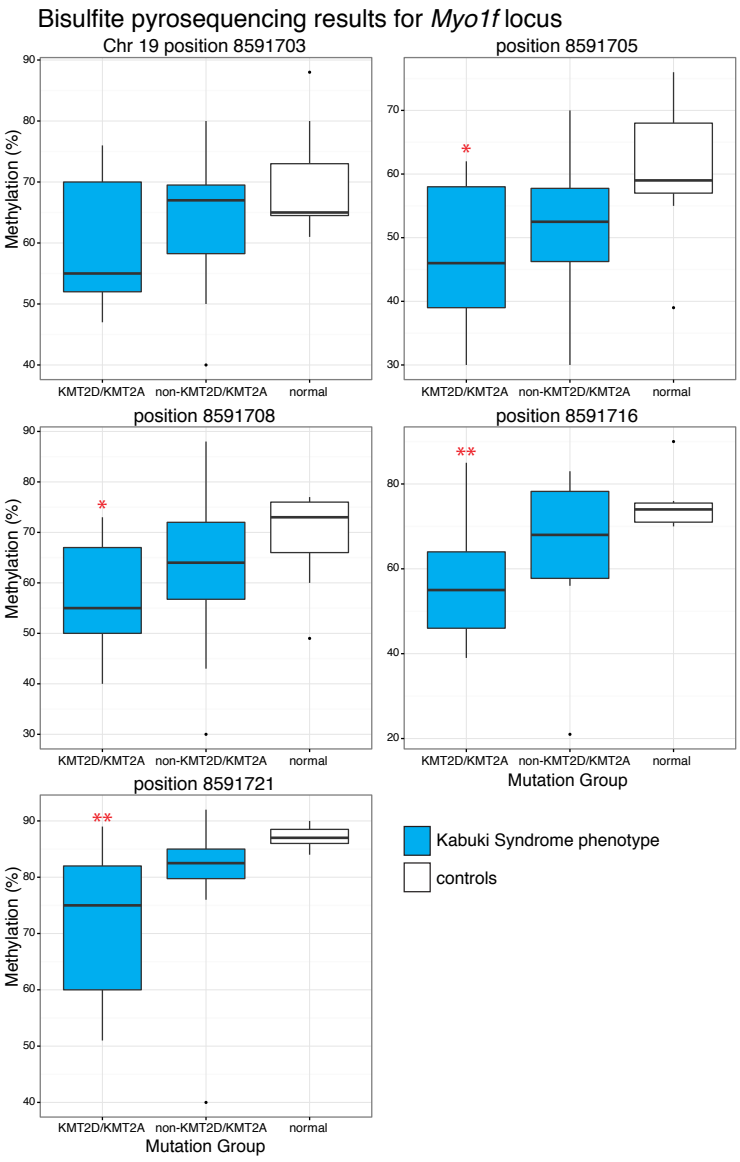
hypomethylation. Similarly, we also identified hypermethylated DMRs in KS. For example, we observed a 20% increase in methylation in KS samples on average, at a DMR located in the 5' UTR region of the *LAMB2* gene (Figure B.7a and Table B.5) which was similarly validated by pyrosequencing in patients with variants in *KMT2D*/ *KMT2A* (Figure B.7b). Again, we observe a similar methylation pattern in patients with a KS phenotype but no discovered variants. These results imply shared downstream targets among individuals with a KS phenotype despite locus heterogeneity for the causative variant.

Figure B.6: *Myo1f* DMR and bisulfite pyrosequencing results

a.



b.

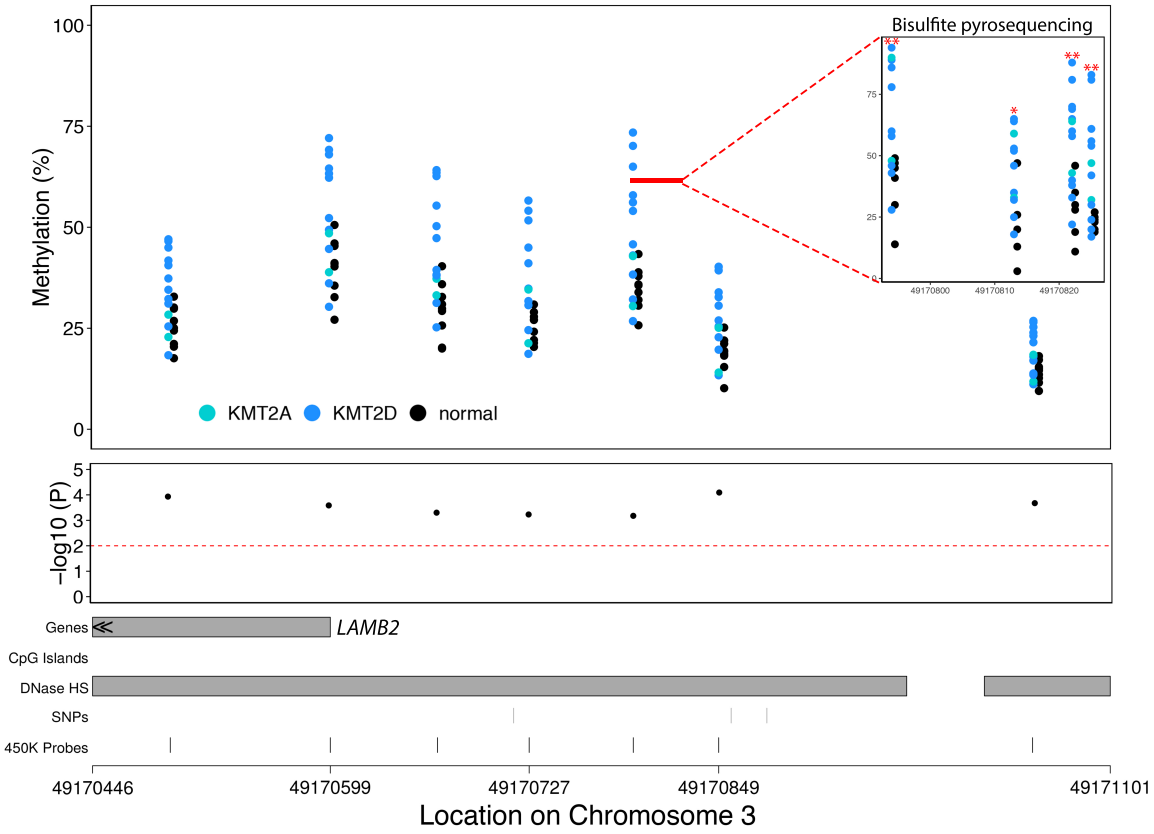


Example of a differentially methylated region (DMR), within the *MYO1F* gene, associated with Kabuki syndrome phenotype and histone methylation machinery gene variants (*KMT2D* or *KMT2A*). A differentially methylated region (DMR), detected via the 450K platform, shows less methylation, on average, among individuals with a histone machinery variant compared to matched individuals without KS (a). The upper

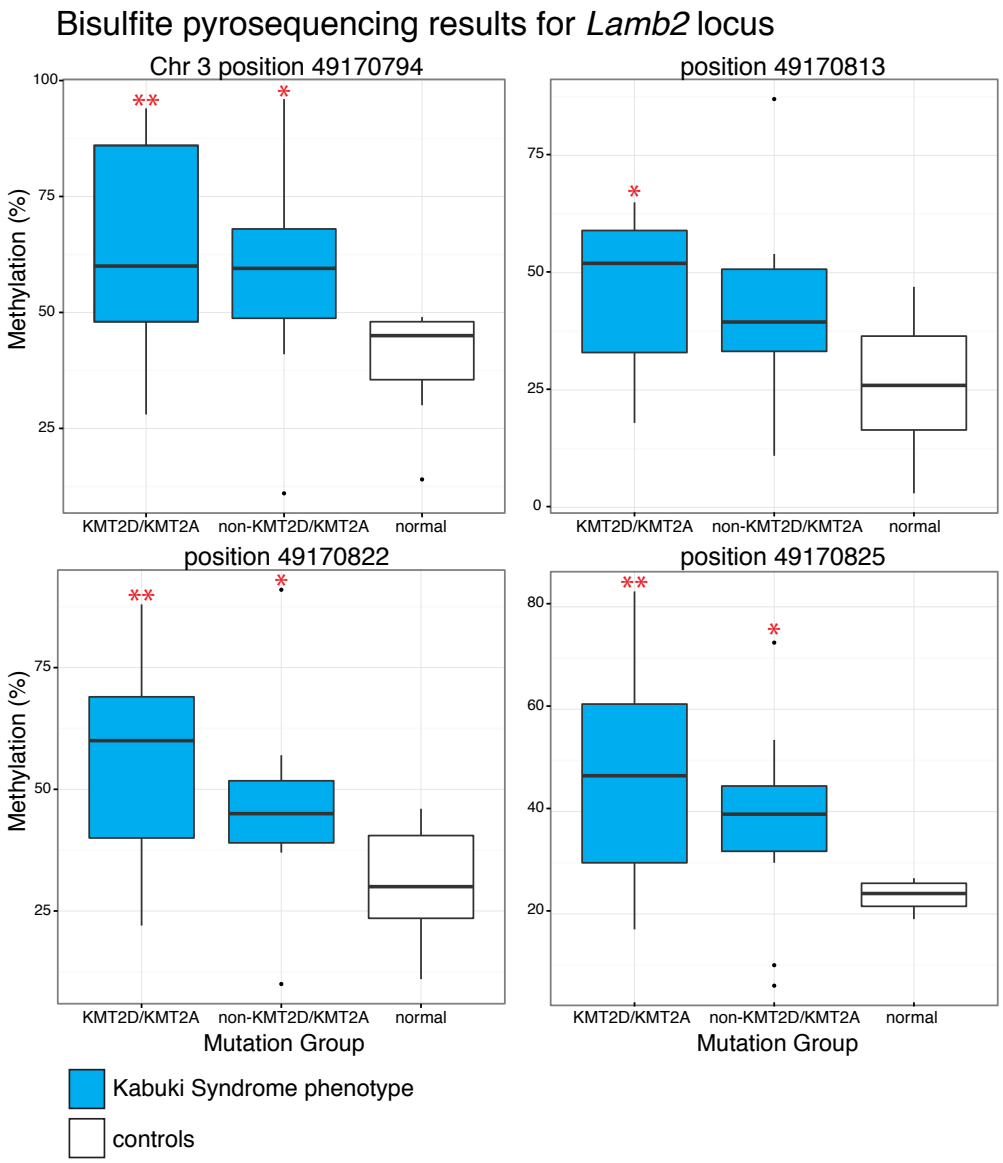
panel displays genomic location on the x-axis and percent methylation on the Y-axis. Inset top panel shows methylation values assessed using bisulfite pyrosequencing for the CpG sites denoted by the solid red line. Blue, green, and black points denote individuals harboring a *KMT2D* variant, a *KMT2A* variant, and controls, respectively. Red asterisks denote significant differential methylation, via bisulfite pyrosequencing, between KS samples with histone machinery variants and control samples. The middle panel shows individual probe-based nominal p-values for differences in methylation between the two groups (KS and control), with the dashed red line at a level of $p = 0.01$. The bottom panel provides gene annotation information for the differentially methylated region. Boxplots comparing bisulfite pyrosequencing derived DNA methylation levels among individuals with histone machinery variants and KS, individuals with no detected genetic variants in our targeted sequencing screen and KS, and control individuals without KS (b). Samples from individuals with KS phenotype are shown in blue and control samples are shown in white. The x-axis denotes the genetic variant group plotted and the y-axis plots the DNA methylation level detected via bisulfite pyrosequencing. Upper and lower hinges of the boxplots correspond to the 75th and 25th percentiles. The upper whisker extends to the highest value within 1.5 times the interquartile range beyond the upper hinge; the lower whisker extends to the lowest value within 1.5 times the interquartile range beyond the lower hinge. Outliers beyond the end of the whiskers are represented as points. *$p<0.05$; **$p<0.01$

Figure B.7: *Lamb2* DMR and bisulfite pyrosequencing results

a.



b.



Relative hypermethylation, at the *LAMB2* locus, associated with Kabuki syndrome phenotype and histone methylation machinery gene variants (*KMT2D* or *KMT2A*) compared to controls. A differentially methylated region (DMR) detected via the 450K platform, shows more methylation, on average, among individuals with a histone machinery variant and KS compared to matched individuals without KS (a). The upper

panel displays genomic location on the x-axis and percent methylation on the Y-axis. Inset top panel shows methylation values assessed using bisulfite pyrosequencing for the CpG sites denoted by the solid red line. Blue, green, and black points denote individuals harboring a *KMT2D* variant, a *KMT2A* variant, and controls, respectively. Red asterisks denote significant differential methylation, via bisulfite pyrosequencing, between KS samples with histone machinery variants and control samples. The middle panel shows individual probe-based nominal p-values for differences in methylation between the two groups (KS and control), with the dashed red line at a level of $p = 0.01$. The bottom panel provides gene annotation information for the differentially methylated region. Boxplots comparing bisulfite pyrosequencing derived DNA methylation levels among individuals with histone machinery variants and KS, individuals with no detected genetic variants in our targeted sequencing screen and KS, and control individuals without KS (b). Samples from individuals with KS phenotype are shown in blue and control samples are shown in white. The x-axis denotes the genetic variant group plotted and the y-axis plots the DNA methylation level detected via bisulfite pyrosequencing. Upper and lower hinges of the boxplots correspond to the 75th and 25th percentiles. The upper whisker extends to the highest value within 1.5 times the interquartile range beyond the upper hinge; the lower whisker extends to the lowest value within 1.5 times the interquartile range beyond the lower hinge. Outliers beyond the end of the whiskers are represented as points. *$p<0.05$; **$p<0.01$

We also performed unsupervised analyses, using hierarchical clustering, to evaluate epigenetic patterns, more globally, using the top 10% most variably methylated probes on the 450K. For the clustering the 482,879 probes that passed QC measures where then restricted to autosomal probes to prevent clustering solely on the basis of patient sex. 11,587 probes on the X and Y chromosome were removed to leave 471,292 autosomal probes. Then we subset to the top 10% of variably methylated probes, leaving 47,130 probes across 22 samples from which a dissimilarity structure was produced using the dist() function in R. Hierarchical clustering was performed with the hclust() function on this dissimilarity structure with the complete linkage method. As shown in Figure B.8, we observed strong clustering of the KS samples with *KMT2D* loss of function variants as well as a cluster of samples primarily comprised of the *KMT2D*/*KMT2A* missense variants. These clusters did not show association with other covariates (Figure B.8 and Table B.6). Thus, DNA methylation patterns, at a more global level appear to be related to the type of histone methylation machinery mutation present in *KMT2D* and *KMT2A*. Although this would support the notion that a Kabuki Syndrome phenotype is associated with a particular epigenotype, it also raises the possibility that specific classes of variants may have unique epigenetic signatures which could act as a biomarker of the disease state; however, further studies will be needed to clarify this.

Although the primary defect in KS is a defect in the histone machinery, it is noteworthy that we also observe site-specific DNA methylation changes relative to controls. DNA methylation abnormalities have also been found in Sotos syndrome (SS, 117550, 614753)[24], another histone methylation machinery disorder. The SS changes

appear more extensive, although the published results do not correct for cell composition or other relevant covariates (sex and ancestry). NSD1, the protein product of the gene most frequently defective in SS, adds both open and closed chromatin modifications (H3K36 and H4K20) whereas *KMT2A/D* only adds open chromatin modifications (H3K4). Therefore, it is possible that the increased number of DNA methylation changes identified in SS compared to KS reflects biologically meaningful differences related to the function of the different histone modifying proteins involved in each disorder. The DNA methylation abnormalities found in SS, WSS and KS suggest interactions between histone and DNA methylation machineries. Additionally, they provide a potential explanation for phenotypic overlap among the Mendelian disorders of the epigenetic machinery[4] (Table B.2).

There has been long standing debate regarding the directionality of epigenetic information flow, i.e. whether DNA methylation dictates histone modification[25] or whether histone modification lead to downstream DNA methylation changes[26,27]. Cells from patients with ICF syndrome (MIM 242860, 614069, 616910, 616911), a condition caused by mutations in a *de novo* DNA methyltransferase and interacting proteins, demonstrate global DNA hypomethylation[28] but also secondary abnormalities of histone modification[28]. Our data show that individuals with *KMT2A/D* mutations also have site-specific DNA methylation changes; thus, providing support for histone modification leading to downstream DNA methylation changes. These results together with those published for SS and ICF suggests that the information flow is bi-directional forming a feedback loop among individual epigenetic machineries. Our data also suggest that DNA

methylation signatures could be diagnostic markers for KS as well as specific types of KS-related variants.

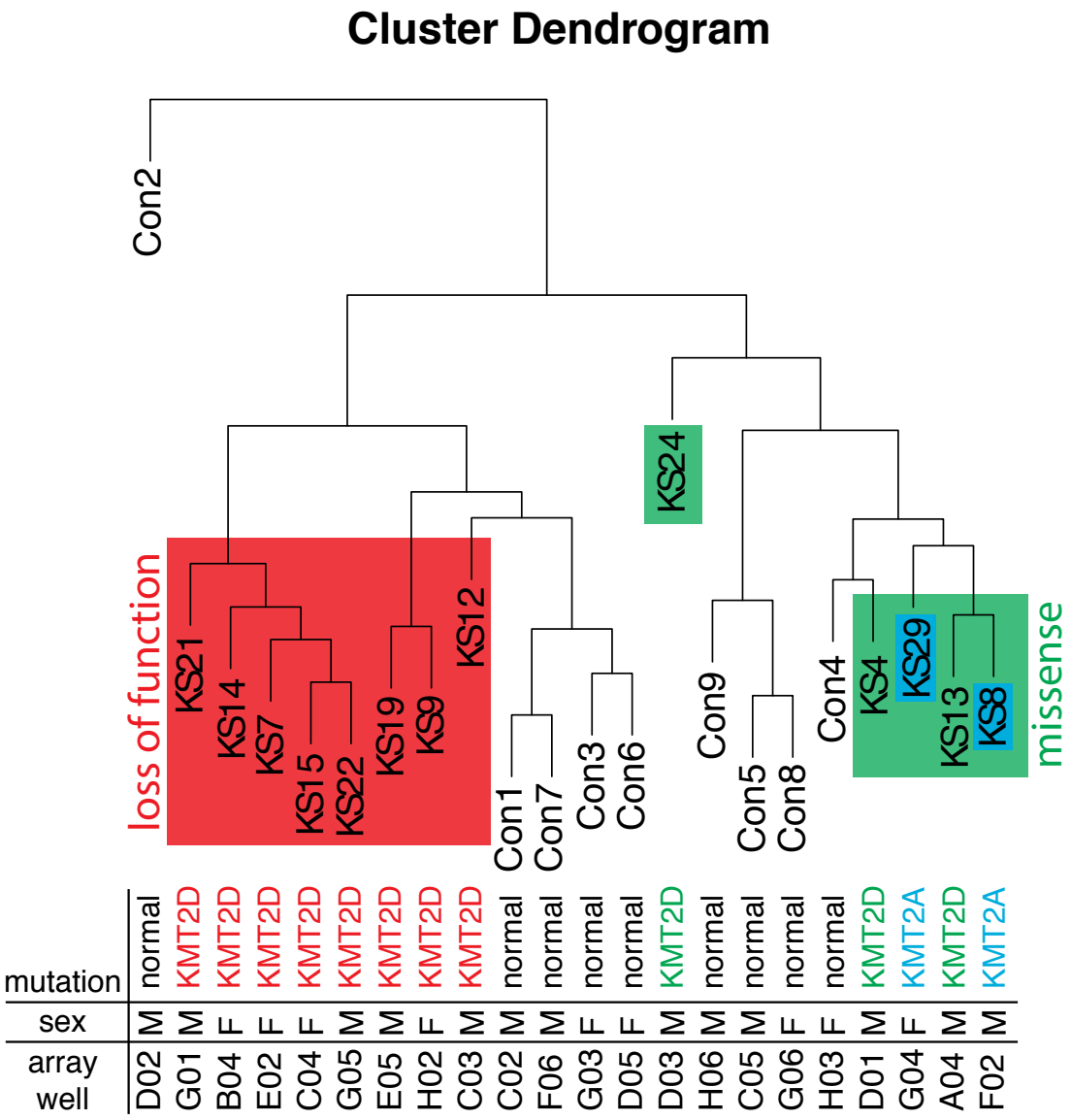


Figure B.8: Hierarchical clustering dendrogram, based on the 10% most variably methylated probes, shows differences in DNA methylation patterns based on type of genetic variation within the histone machinery genes *KMT2A* and *KMT2D*.

Here we depict loss of function *KMT2D* variants (red) and *KMT2D* missense changes (green). Interestingly, the *KMT2A* missense variants (blue) cluster with the

KMT2D missense variants, indicating that epigenetic similarities may account for phenotypic overlap.

Table B.6: Estimated amount of each cell type. No patterns emerge to suggest cell type composition drives clustering

	CD8 T-cells	CD4 T-cells	NK	B-cells	Monocytes	Granulocytes
Con1	0.100	0.136	0.034	0.097	0.026	0.623
Con2	0.103	0.113	0.011	0.055	0.228	0.519
Con3	0.042	0.117	0.088	0.026	0.125	0.595
Con4	0.145	0.245	0.000	0.088	0.050	0.478
Con5	0.200	0.228	0.009	0.156	0.013	0.399
Con6	0.072	0.168	0.097	0.059	0.031	0.578
Con7	0.094	0.187	0.032	0.047	0.052	0.596
Con8	0.180	0.214	0.022	0.198	0.039	0.339
Con9	0.144	0.136	0.073	0.189	0.059	0.427
PAT12	0.103	0.065	0.000	0.105	0.088	0.635
PAT13	0.213	0.091	0.042	0.121	0.103	0.433
PAT14	0.016	0.114	0.000	0.235	0.151	0.474
PAT15	0.094	0.237	0.000	0.147	0.064	0.449
PAT19	0.101	0.082	0.001	0.062	0.077	0.679
PAT21	0.157	0.122	0.000	0.095	0.086	0.529
PAT22	0.128	0.218	0.000	0.063	0.045	0.536
PAT24	0.044	0.225	0.032	0.131	0.109	0.438
PAT29	0.072	0.150	0.223	0.040	0.059	0.424
PAT4	0.142	0.143	0.136	0.097	0.097	0.402
PAT7	0.092	0.194	0.034	0.178	0.073	0.429
PAT8	0.097	0.231	0.019	0.160	0.069	0.422
PAT9	0.038	0.137	0.000	0.057	0.092	0.668

Acknowledgements

We would like to thank all the families that participated in this study. We also thank Maggie Baker for assistance in selecting the candidate genes for the amplicon study. This work was supported by a grant to H.T.B. by the NIH Director's Early Independence Award (DP5OD017877) and a grant to D.V. from the National Human Genome Research Institute (1U54HG006493).

Web resources

The URLs for data presented herein are as follows:

Baylor-Hopkins Center for Mendelian Genomics, http://mendeliangenomics.org;

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA, http://evs.gs.washington.edu/EVS/;

1000 Genomes, http://www.1000genomes.org;

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/;

ExAC browser, http://exac.broadinstitute.org/;

Picard software, http://picard.sourceforge.net;

PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/ibdibs.shtml;

PolyPhen-2, http://www.genetics.bwh.harvard.edu/pph2/;

SIFT, http://sift.bii.a-star.edu.sg/

InterPro, https://www.ebi.ac.uk/interpro/

References

1. Ng S.B., Bigham A.W., Buckingham K.J., Hannibal M.C., McMillin M.J., Gildersleeve H.I., Beck A.E., Tabor H.K., Cooper G.M., Mefford H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. *42*, 790-3.
2. Miyake N., Mizuno S., Okamoto N., Ohashi H., Shiina M., Ogata K., Tsurusaki Y., Nakashima M., Saitsu H., Niikawa N., et al. (2013). KDM6A point mutations cause Kabuki syndrome. Hum Mutat. *34*, 108-10.
3. Lindsley A.W., Saal H.M., Burrow T.A., Hopkin R.J., Shchelochkov O., Khandelwal P., Xie C., Bleesing J., Filipovich L., Risma K., et al. (2016). Defects of B-cell terminal differentiation in patients with type-1 Kabuki syndrome. J Allergy Clin Immunol. *137*, 179-87.
4. Bjornsson H.T. (2015). The Mendelian disorders of the epigenetic machinery. Genome Res. *25*, 1473-81.
5. Jones W.D., Dafou D., McEntagart M., Woollard W.J., Elmslie F.V., Holder-Espinasse M., Irving M., Saggar A.K., Smithson S., Trembath R.C., et al. (2012) De novo mutations in MLL cause Wiedemann-Steiner syndrome. Am J Hum Genet. 91, 358-64.
6. Hamosh, A., Sobreira N., Hoover-Fong J., Sutton V.R., Boehm C., Schiettecatte F., Valle D. (2013). PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. Hum Mutat. 34, 566-71.

7. Adam M.P., Hudgins L. (2005). Kabuki syndrome: a review. Clin Genet. *67*, 209-19
8. Dentici M.L., Di Pede A., Lepri F.R., Gnazzo M., Lombardi M.H., Auriti C., Petrocchi S., Pisaneschi E., Bellacchio E., Capolino R., et al. Kabuki syndrome: clinical and molecular diagnosis in the first year of life. Arch Dis Child. *100*, 158-64.
9. Sobreira N., Schiettecatte F., Boehm C., Valle D., Hamosh A. (2015). New tools for Mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. Hum Mutat. 36, 425-31.
10. The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.
11. Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G., Mesirov J.P. (2011). Integrative genomics viewer. Nat Biotechnol. 29, 24-6.
12. Mendelsohn B.A., Pronold M., Long R., Smaoui N., Slavotinek A.M. (2014). Advanced bone age in a girl with Wiedemann-Steiner syndrome and an exonic deletion in KMT2A (MLL). Am J Med Genet A. 8, 2079-83.
13. Stellacci E., Onesimo R., Bruselles A., Pizzi S., Battaglia D., Leoni C., Zampino G., Tartaglia M. (2016). Congenital immunodeficiency in an individual with Wiedemann-Steiner syndrome due to a novel missense mutation in KMT2A. Am J Med Genet A. 170, 2389-93.

14. Bögershausen N., Gatinois V., Riehmer V., Kayserili H., Becker J., Thoenes M., Simsek-Kiper P.Ö., Barat-Houari M., Elcioglu N.H., Wieczorek D., et al. (2016). Mutation Update for Kabuki Syndrome Genes KMT2D and KDM6A and Further Delineation of X-Linked Kabuki Syndrome Subtype 2. Hum Mutat. *37*, 847-64.
15. Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.
16. Aryee M.J., Jaffe A.E., Corrada-Bravo H., Ladd-Acosta C., Feinberg A.P., Hansen K.D., Irizarry R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 30, 1363-9.
17. Du P., Zhang X., Huang C.C., Jafari N., Kibbe W.A., Hou L., Lin S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. *11*, 587.
18. Houseman E.A., Accomando W.P., Koestler D.C., Christensen B.C., Marsit C.J., Nelson H.H., Wiencke J.K., Kelsey K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 13, 86.
19. Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., Smyth G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
20. Barfield R.T., Almli L.M., Kilaru V., Smith A.K., Mercer K.B., Duncan R., Klengel T., Mehta D., Binder E.B., Epstein M.P., et al. (2014). Accounting for

population stratification in DNA methylation studies. Genet Epidemiol. 38, 231-41.

21. Jaffe A.E., Murakami P., Lee H., Leek J.T., Fallin M.D., Feinberg A.P., Irizarry R.A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. *41*, 200-9.
22. Li L.C., Dahiya R. (2002). MethPrimer: designing primers for methylation PCRs. Bioinformatics. 18, 1427-31.
23. Tost J. Gut I. G. (2007). DNA methylation analysis by pyrosequencing. Nature Protocols. 2, 2265-2275.
24. Choufani S., Cytrynbaum C., Chung B.H., Turinsky A.L., Grafodatskaya D., Chen Y.A., Cohen A.S., Dupuis L., Butcher D.T., Siu M.T., et al (2015). NSD1 mutations generate a genome-wide DNA methylation signature. Nat Commun. 6, 10207.
25. Okitsu C.Y., Hsieh C.L. (2007). DNA methylation dictates histone H3K4 methylation. Mol Cell Biol. *27*, 2746-57.
26. Hu J.L., Zhou B.O., Zhang R.R., Zhang K.L., Zhou J.Q., Xu G.L. (2009). The N-terminus of histone H3 is required for de novo DNA methylation in chromatin. Proc Natl Acad Sci U S A. *106*, 22187-92.
27. Wang J., Hevi S., Kurash J.K., Lei H., Gay F., Bajko J., Su H., Sun W., Chang H., Xu G., et al. (2009) The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation. *41*, 125-9.

28. Jin B., Tao Q., Peng J., Soo H.M., Wu W., Ying J., Fields C.R., Delmas A.L., Liu X., Qiu J., Robertson K.D. (2008). DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. Hum Mol Genet. *17*, 690-709.

Curriculum Vitae

Martha Frances Brucato
born on January 10, 1987 in Alexandria, Virginia
married October 22, 2013 to Benjamin Michael O'Neil in Baltimore, Maryland

18 N. Luzerne Avenue
Baltimore, MD 21224
(919) 943-3519
mbrucat1@jhmi.edu or martha.brucato@gmail.com

EDUCATION

Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
Department of Epidemiology
PhD candidate in Genetic Epidemiology
Entered Fall 2013

Johns Hopkins University School of Medicine
Medical Scientist Training Program, entered 2010
MD/PhD expected 2018

Duke University, Durham, North Carolina
Bachelor of Science, *summa cum laude*, May 2009
Majors: Biology, Asian & Middle Eastern Studies (Japanese concentration)
Minor: Chemistry

RESEARCH EXPERIENCE

Johns Hopkins University Bloomberg School of Public Health
Doctoral Student (September 2013–Present)
Thesis Advisor: Christine Ladd-Acosta, PhD
utilizing administrative health data for epidemiologic research on complex phenotypes like Autism Spectrum Disorders (ASD)
environmental epigenetic epidemiology in complex disease (e.g. ASD) and Mendelian genetic disease (e.g. Kabuki Syndrome)

Johns Hopkins University School of Medicine
Rotation Student (June 2011–December 2011)
Principal Investigator: Andrew Ewald, PhD
organotypic culture of mammary gland organoids

developed whole-organ model system for tracking *ex vivo* development of mammary gland ducts

Johns Hopkins University School of Medicine
Rotation Student (Summer 2010)
Principal Investigator: Elizabeth Jaffee, MD
knocked down mesothelin, a protein known primarily as a tumor antigen and suspected to be involved in metastasis, in mouse pancreatic tumor cells
learned cell culture, flow cytometry, cell sorting, mouse spleen dissection, lentivirus production and infection

Duke University Eye Center
Research Technician (October 2006–May 2010)
Principal Investigator: Vadim Arshavsky, PhD
researched the intracellular trafficking of the proteins retinal guanylate cyclase 1, peripherin2-RDS, synaptophysin, and others in photoreceptors
joined the lab full-time post-graduation

NIH/National Institute of Neurological Disorders and Stroke,
Bethesda, Maryland
Biological Science Laboratory Technician (Summer 2006, 2007, 2008)
Principal Investigator: Heinz Arnheiter, MD
investigated the function of Microphthalmia Associated Transcription Factor (Mitf) by analyzing the retinal and retinal pigment epithelium phenotypes in several mutant mice, including compound mutants of Mitf and Chx10
studied the interactions of Pax6 and Mitf, investigating the mechanistic causes of the changes found in Pax6/Mitf compound mutants

Duke University Medical Center, Obstetric Anesthesiology
Research Technician (September 2005–May 2008)
Principal Investigator: James Reynolds, PhD
investigated the effect of ethyl nitrite gas on hypoxic pregnant sheep
researched the impact of long-term carbon insufflation on body tissues in pigs
studied the effect of MDMA ("ecstasy") on pregnant sheep and their fetuses

PUBLICATIONS

Bharti, K., Gasper, M., Ou, J., Brucato, M., Clore-Gronenberg, K., Pickel, J., Arnheiter, H. "A regulatory loop involving PAX6, MITF, and WNT signaling controls retinal pigment epithelium development." PLoS Genet. 2012 Jul; 8(7):e1002757. doi: 10.1371/journal.pgen.1002757.

Gospe, S., Baker, S., Kessler, C., Brucato, M., Winter, J., Burns, M., Arshavsky, V. "Membrane attachment is key to protecting transducin GTPase activating complex from intracellular proteolysis in photoreceptors." J Neuroscience 2011 Oct 12;31(41): 14660-8.

ABSTRACTS

Brucato, M., Ladd-Acosta, C., Li, M., Caruso, D., Hong, X., Wang, X., Fallin, D. (2016). The association between prenatal exposure to maternal infection and Autism Spectrum Disorder in the Boston Birth Cohort. Poster. International Meeting for Autism Research 2016, Baltimore, Maryland.

Brucato, M., Sobreira, N., Zhang, L., Ladd-Acosta, C., Ongaco, C., Romm, J., Baker, M., Doheny, K., Bertola, D., Chong, K., Perez, A.B.A., Melaragno, M., Meloni, V., Valle, D., and Bjornsson, H.T. (2015). Comparison of Illumina Infinium 450K Methylation BeadChip preprocessing methods in an Epigenome Wide Association Study. Poster. International Genetic Epidemiology Society 2015 Annual Meeting, Baltimore, Maryland.

Brucato, M., Sobreira, N., Zhang, L., Ladd-Acosta, C., Ongaco, C., Romm, J., Baker, M., Doheny, K., Bertola, D., Chong, K., Perez, A.B.A., Melaragno, M., Meloni, V., Valle, D., and Bjornsson, H.T. (2015). Kabuki syndrome, a disorder of histone methylation, demonstrates characteristic DNA methylation abnormalities illustrating potential interplay between histone and DNA methylation machineries. Poster. MD-GEM Genetics Research Day, Baltimore, Maryland.

Brucato, M., Arshavsky, V.Y., and Baker, S.A. (2010). Identification of a Targeting Signal in the Synaptic Vesicle Protein, Synaptophysin. Poster. Association for Research in Vision and Ophthalmology 2010 Annual Meeting, Ft. Lauderdale, Florida.

Bharti, K., Gasper, M., Brucato, M., Maminishki, A., Miller, S.S., and Arnheiter, H. (2010). PAX6 and MITF Play a Dose Dependent Role in Determining Cell Fate of the Retinal Pigment Epithelium (RPE) and Putative Ocular Stem Cells. Oral Presentation. Association for Research in Vision and Ophthalmology 2010 Annual Meeting, Ft. Lauderdale, Florida.

Brucato, M., Arshavsky, V.Y., and S.A. Baker. (2009). Identification of a Targeting Signal in the Synaptic Vesicle Protein, Synaptophysin. Oral Presentation. American Society for Cell Biology conference, San Diego, California.

AWARDS/HONORS

Wendy Klag Center Student Award, "Developing methods utilizing Machine Learning and Latent Class Analysis to identify children with ASD in administrative health data" (May 2016)

MD-GEM Wolfe Street Competition prize for the proposal, "Detection of novel genetic mechanisms for strabismus through whole genome sequencing" (February 2016)

Johns Hopkins MD-GEMS Genetics Research Day poster competition, first place among doctoral students (February 2015)

Baltimore Albert Schweitzer Fellowship to address health disparities in Baltimore and develop leadership in community service (April 2012)

Excellence in Medical Student Research award winner for "Community Adolescent Sexuality Education: Program Needs Assessment and Curriculum Update" (March 2012)

Johns Hopkins Alumni Association Grant for Community Adolescent Sexuality Education (CASE) project (Dec 2011)

Selection to lecture to incoming JHUSOM students on Health Care Disparities (Aug 2011)

Assist development of the Henrietta Lacks Scholarship, created by JHHS and JHUSOM

Medical Students for Choice Travel Grant to the 2010 Annual Meeting (Dec 2010)

National Eye Institute Travel Grant to the ARVO 2010 Annual Meeting (May 2010)

Dean's List (2005–2009) with Distinction (Fall 2005, Fall 2006–Spring 2009)

Alice M. Baldwin Scholarship, Duke University named scholarship (2008–2009)

Dean's Summer Research Fellowship (2008)

Asian/Pacific Studies Institute at Duke University Research Grant (2008)

Phi Beta Kappa (Spring 2008)

LEADERSHIP AND SERVICE EXPERIENCES

Preceptor for First Year Medical Students with the Student Preceptor Program (August 2013–May 2014)

Baltimore-Kawasaki Sister City Committee member (March 2012–2014)

Community Adolescent Sexuality Education, coordinator and teacher (2010–2014)

Student National Medical Association, member (2010–present) and chapter President (2011–2012)

Medical Students for Choice, Johns Hopkins Chapter Founder and President (2010–2011)

Healthy Devils Peer Educator: Duke Educational Leaders in Sexual Health (2006–2009)

Organic Chemistry Peer Tutor (2006–2009)

Resident Assistant (2006–2009)

Internship at the Center for the Health and Rights of Migrants in Osaka, Japan (2008)

Volunteer Patient Advocate at Duke Emergency Department (2007–2008)

ATHLETICS

Powerlifting

USAPL Terrapin Open, 2/25/2017
Squat 92.5kg/204lbs
Bench 52.5kg/116lbs
Deadlift 107.5kg/237lbs
at 60.5 kg bodyweight